Modelling Australian annual mean rainfall data: a new approach based on fractional integration

L.A. Gil-Alana

University of Navarra, Pamplona, Spain

This paper analyses Australian annual rainfall data from 1900 to 2006. We use new statistical techniques based on the concept of fractional integration, which allow us to examine the dynamic behaviour of the series in a much richer way than the classic approaches based on stationary or nonstationary processes. Moreover, the possibility of a structural break at an unknown date is also taken into account. The results indicate that a break takes place at 1973, the time trend coefficients being insignificant in the two subsamples, and the degree of dependence between the observations is higher after the break. Regional data are also explored and the results show significant differences in the degree of persistence across the regions, especially after the breaks.

Introduction

The modelling of rainfall and other hydrological data has been the point of much discussion in past research. Hurst (1951) heuristically detected the presence of long range dependence (or long memory) in the well-known series of annual minima of the Nile River. These processes are so called because they display a high degree of association between observations which are distant in time. Following Hurst's work, extensive research has been carried out to detect long memory in hydrological data. According to Montanari and Rosso (1997) there are two main reasons for the presence of the Hurst phenomenon in time series: (a) the series exhibits persistence, i.e. a strong association between observations widely separated in time, and/or (b) the mean of the series changes with time, i.e. the series is nonstationary. Nowadays, it is widely accepted that the Hurst phenomenon may be caused by both serial correlation and nonstationarity. In another study, Montanari et al. (1996) examined six rainfall time series in various sites in Italy, in order to analyse whether a linear trend and/or long memory is present in the data. Among other methods, they propose a modified

variance-type estimator to detect long memory even when some types of nonstationarity are potentially present. They found a decreasing trend, though not statistically significant, in each of their records, but significant long memory in only two of them.

In this article we examine the long memory property of annual Australian-averaged rainfall data using fractional integration techniques. By fractional integration we mean that the order of differencing required to get a series that is stationary may be a fractional value. A proper discussion of this idea and other relevant statistical concepts will be presented in the following section. We employ a procedure that permits us to estimate the fractional differencing parameter even in nonstationary contexts. Moreover, the presence of a changing deterministic pattern is also examined through a method that allows us to estimate the deterministic terms and the fractional orders of integration at each subsample, the break-date being endogenously determined by the model itself.

There are numerous papers describing analysis of Australian rainfall data. For example, McBride and Nicholls (1983) showed that the variability of rainfall across Australia is strongly influenced by the El Niño-Southern Oscillation (ENSO) on interannual time scales. Chiew and McMahon (2003) also demonstrated a link between ENSO and Australian rainfall. More recently, Srikanthan et al. (2007) argued that long-term persistence in Australian annual rainfall data

Corresponding author address:
Luis A. Gil-Alana, University of Navarra, Faculty of Economics, Edificio Biblioteca, Entrada Este, E-31080 Pamplona, Spain.
E-mail: alana@unav.es.

results from climatic fluctuations between wet and dry periods influenced by the ENSO phenomenon and the inter-decadal Pacific Oscillation. When exploring Australian rainfall on a regional scale, Simmonds and Hope (1997) identified statistically significant persistence on monthly, seasonal and annual time scales, together with significant correlation with the SOI (Southern Oscillation Index). Another recent study modelling persistence in annual Australian point rainfall data is described in Whiting et al. (2003). These authors examined annual rainfall time series for Sydney from 1859 to 1999 and showed clear evidence of nonstationarity in the data.

Terminology and statistical concepts

A time series process $\{x_t, t = 0, \pm 1, \ldots\}$ is said to be (covariance) stationary if the mean and the variance do not depend on time and the covariance between any two observations depends on the temporal distance between them but not on their specific location in time. This is a minimal requirement in time series analysis to make statistical inference.

Given a zero-mean covariance-stationary process $\{x_t, t = 0, \pm 1, \ldots\}$, with autocovariance function $\gamma_u = E(x_t, x_{t+u})$, we say that x_t is integrated of order zero (denoted by $x_t \approx I(0)$) if

$$\sum_{u=-\infty}^{\infty} |\gamma_u| < \infty .$$

Examples of $I(0)$ processes are the familiar white noise, stationary autoregressions (AR), moving averages (MA), autoregressive moving averages (ARMA), etc. If the time series is nonstationary, one possibility for transforming the series into a stationary one is to take first differences, such that

$$(1 - L)\, x_t = u_t, \qquad t = 1, 2, \ldots \qquad \ldots(1)$$

where L is the lag-operator ($L x_t = x_{t-1}$) and u_t is $I(0)$ as defined above. In such a case, x_t is said to be integrated of order 1 (denoted by $x_t \approx I(1)$). Likewise, if two differences are required, the series is integrated of order 2 ($I(2)$). If the number of differences required to get $I(0)$ stationarity is not an integer value but a fractional one, the process is said to be fractionally integrated or $I(d)$. In other words, we say that x_t is $I(d)$ if

$$(1 - L)^d x_t = u_t, \qquad t = 1, 2, \ldots \qquad \ldots(2)$$

with $I(0)$ u_t. Note that the expression in the left-hand-side in Eqn 2 can be presented in terms of its Binomial expansion, such that, for all real d,

$$(1 - L)^d = \sum_{j=0}^{\infty} \psi_j L^j = \sum_{j=0}^{\infty} \binom{d}{j} (-1)^j L^j$$

$$= 1 - d\,L + \frac{d\,(d-1)}{2} L^2 - \ldots,$$

(Beran 1994), and Eqn 2 can be written as:

$$x_t = d\,x_{t-1} - \frac{d\,(d-1)}{2} x_{t-2} + \ldots + u_t$$

If d is a positive integer value, x_t will be a function of a finite number of past observations, while if d is not an integer, x_t depends strongly upon values of the time series far in the past (e.g., Granger and Ding 1996; Dueker and Asea 1998). Moreover, the higher the value of d, the higher will be the level of association between the observations.

The parameter d plays a crucial role from a statistical viewpoint. Thus, if $d < 0$, the series is said to be anti-persistent (Mandelbrot 1977); if $d = 0$, x_t is stationary $I(0)$; if $0 < d < \frac{1}{2}$, the series is fractionally integrated though still covariance stationary, and as d increases beyond $\frac{1}{2}$ and through 1 (the unit root case) x_t can be viewed as becoming "more nonstationary" in the sense, for example, that the variance of the partial sums increases in magnitude. This is also true for $d > 1$, so a large class of nonstationary processes may be described by Eqn 2 with $d \geq \frac{1}{2}$.

The statistical model

We seek to model the observed time series y_t as the sum of a linear trend and a fractionally integrated noise term over two subsamples

$$y_t = \alpha_1 + \beta_1 t + x_t;$$
$$(1 - L)^{d_1} x_t = u_t, \qquad t = 1, \ldots, T_b \qquad \ldots(3)$$

$$y_t = \alpha_2 + \beta_2 t + x_t;$$
$$(1 - L)^{d_2} x_t = u_t, \qquad t = T_b + 1, \ldots, T, \qquad \ldots(4)$$

where the αs and the βs are the coefficients corresponding respectively to the intercepts and the linear trends; d_1 and d_2 may be real values, u_t is $I(0)$ and T_b is the location of an unknown break to be estimated from the data. Note that the model in Eqns 3 and 4 can also be written as

$$(1 - L)^{d_1} y_t = \alpha_1 \tilde{1}_t(d_1) + \beta_1 \tilde{t}_t(d_1) + u_t,$$
$$t = 1, \ldots, T_b, \qquad \ldots(5)$$

$$(1 - L)^{d_2} y_t = \alpha_2 \tilde{1}_t(d_2) + \beta_2 \tilde{t}_t(d_2) + u_t,$$
$$t = T_{b+1}, \ldots, T \qquad \ldots(6)$$

where $\tilde{1}_t(d_i) = (1 - L)^{d_i} 1$

and $\tilde{t}_t(d_i) = (1 - L)^{d_i} t$ (for $i = 1, 2$).

The method presented here is based on the least squares principle and is similar to the one employed in Bai and Perron (1998).[1] First we choose a grid for the values of the fractional differencing parameters d_1 and d_2, for example, $d_{io} = -1, -0.99, -0.98, \ldots, 1.99, 2$ ($i = 1, 2$). Then, for a given partition with a break location at $t = \{T_b\}$ and given d_1, d_2 values (d_{1o}, d_{2o}), we estimate the αs and the βs by minimizing the sum of squared residuals,

$$\min\left\{\sum_{t=1}^{T_b}\left[(1-L)^{d_{1o}}y_t - \alpha_1\,\tilde{1}_t(d_{1o}) - \beta_1\,\tilde{t}_t(d_{1o})^2\right] + \right.$$
$$\left. + \sum_{t=T_b+1}^{T}\left[(1-L)^{d_{2o}}y_t - \alpha_2\,\tilde{1}_t(d_{2o}) - \beta_2\,\tilde{t}_t(d_{2o})^2\right]\right\}$$

$$w.r.t.\{\alpha_1, \alpha_2, \beta_1, \beta_2\}$$

in case of uncorrelated disturbances or using Generalized Least Squares (GLS) with autocorrelated u_t.

Let $\hat{\alpha}(T_b; d_{1o}^{(1)}, d_{2o}^{(1)})$ and $\hat{\beta}(T_b; d_{1o}^{(1)}, d_{2o}^{(1)})$

denote the resulting estimates for a partition at $t = \{T_b\}$ and initial values $d_{1o}^{(1)}$ and $d_{2o}^{(1)}$. Substituting these estimated values on the objective function in Eqns 3 and 4, we have RSS(T_b; $d_{1o}^{(1)}$, $d_{2o}^{(1)}$), and minimizing this expression across all values of d_{1o} and d_{2o} in the grid we obtain

$$RSS(T_b) = \min_{\{m,n\}} RSS(T_b; d_{1o}^{(m)}, d_{2o}^{(n)}),$$

where m and n are the number of values in the grid-search. Then, the estimated break point, $\hat{T}_k$, is such that

$$\hat{T}_k = \min_{i=\{T_1, T_2, \ldots, T_m\}} RSS(T_i),$$

where the minimization is taken over all partitions T_1, T_2, …, T_m, such that $T_i - T_{i-1} \geq |\varepsilon T|$.[2] Then, the regression parameter estimates are the associated least-squares estimates of the estimated k-partition, i.e., $\hat{\alpha}_i = \hat{\alpha}_i(\{\hat{T}_k\}), \hat{\beta}_i = \hat{\beta}_i(\{\hat{T}_k\})$, and their corresponding differencing parameters, $\hat{d}_i = \hat{d}_i(\{\hat{T}_k\})$, for $i = 1$ and 2. Several Monte Carlo results based on the model in Eqns 3 and 4 are provided in Gil-Alana (2008). In that paper, the author shows that the method performs relatively well even with samples as small as those employed in the present work.

The Australian rainfall data

Australia's annual mean rainfall is an area-weighted aver-

age of the total annual precipitation at approximately 370 high-quality rainfall stations around the country. The rainfall records of these stations form part of a dataset developed to monitor long-term trends in Australian rainfall. These data are obtained from the Australian Bureau of Meteorology, and they are available at http://www.bom.gov.au/silo/products/cli_chg/ (see also Lavery et al., 1992, 1997). The data cover the period 1900-2006.

The upper part of Fig. 1 displays a plot of the rainfall series. A slight increase in Australian annual mean rainfall is evident during the 20th century although this is largely due to several wet years during the 1970s. However, the high year-to-year variability of Australian rainfall dominates any background trends. Some of this variability can be accounted for by the El Niño-Southern Oscillation. The second plot in the figure displays the sample autocorrelation values. We observe various significant values even at some large lags suggesting that long memory may be present in the data. The periodogram (displayed in the lower part of the figure)

Fig.1　(a) Annual mean rainfall data in Australia. (b) Correlogram of the original data. (c) Periodogram of the original data. The large sample standard error under the null hypothesis of no autocorrelation is $1/\sqrt{T}$ or roughly 0.096. The periodogram was computed based on the discrete Fourier frequencies $\lambda_j = 2\pi j/T$. T is the sample size, i.e. 107 observations.

---
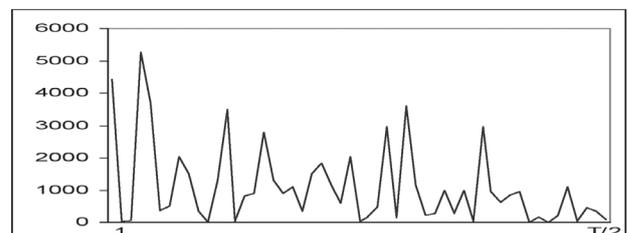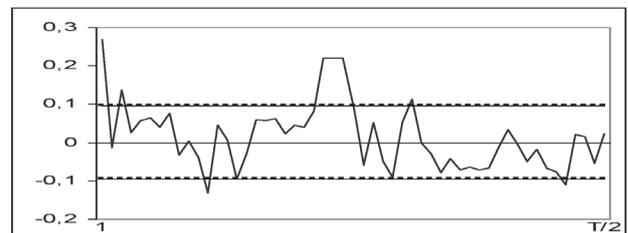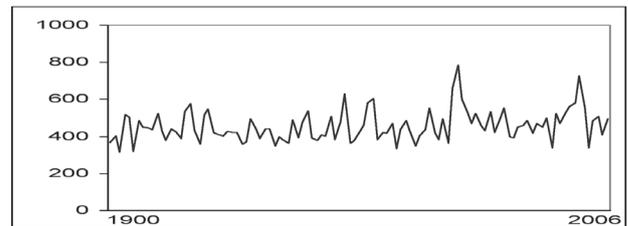
[1]The method proposed in Bai and Perron (1998) is valid for I(0) disturbance errors. Gil-Alana (2008) extends that approach to the case of I(d) processes where d can be any real value.

[2]This means that we consider in the grid-search all potential values except some extreme cases. This is standard in the structural change literature (e.g. Bai and Perron, 1998; Banerjee and Urga 2005).

shows peaks at zero and at various other frequencies. The periodogram is an asymptotically unbiased estimator of the spectral density function. If a series is $I(d)$ with $d > 0$ the spectrum is unbounded at a given frequency, and the periodogram should reproduce that behaviour.

In this study we also employ data corresponding to six climatologically distinct regions; northern Australia (north of 26°S); southern Australia (south of 26°S); southwestern Australia (southwest of the line joining 30°S, 115°E and 35°S, 120°E); southeastern Australia (south of 33°S, east of 135°E); eastern Australia (Queensland, New South Wales, Victoria and Tasmania), and the Murray-Darling Basin. Values calculated for New South Wales include the Australian Capital Territory.

Empirical results

The first thing we do is estimate the fractional differencing parameter assuming that there are no breaks in the data. For this purpose we employ a parametric approach suggested by Robinson (1994). We test the null hypothesis

$$H_0 : d = d_o, \qquad \qquad ...(7)$$

in a model given by

$$y_t = \alpha + \beta t + x_t, \qquad t = 1, 2, ... \quad ...(8)$$

with x_t given by

$$(1 - L)^d x_t = u_t, \qquad \qquad ...(9)$$

and $I(0)u_t$. This method is based on the Lagrange Multiplier (LM) principle and tests H_0 (7) for any real value d_o, including thus the cases of $I(0)$ stationarity ($d_o = 0$, i.e., $x_t = u_t$ in Eqn 9), nonstationarity ($d_o \geq \frac{1}{2}$), and the unit root case when $d_o = 1$. The test statistic takes the form

$$\hat{r} = \frac{T^{1/2}}{\hat{\sigma}^2} \hat{A}^{-1/2} \hat{a} \qquad \qquad ..(10)$$

where T is the sample size and

$$\hat{a} = \frac{-2\pi}{T} \sum_{j=1}^{T-1} \psi(\lambda_j) g(\lambda_j; \hat{\tau})^{-1} I(\lambda_j);$$

$$\hat{\sigma}^2 = \sigma^2(\hat{\tau}) = \frac{2\pi}{T} \sum_{j=1}^{T-1} g(\lambda_j; \hat{\tau})^{-1} I(\lambda_j);$$

$$\hat{A} = \frac{2}{T} \left(\sum_{j=1}^{T-1} \psi(\lambda_j)^2 - \sum_{j=1}^{T-1} \psi(\lambda_j) \hat{\varepsilon}(\lambda_j)' \right.$$

$$\left. \times \left(\sum_{j=1}^{T-1} \hat{\varepsilon}(\lambda_j) \hat{\varepsilon}(\lambda_j)' \right)^{-1} \times \sum_{j=1}^{T-1} \hat{\varepsilon}(\lambda_j) \psi(\lambda_j) \right)$$

$$\psi(\lambda_j) = \log \left| 2 \sin(\frac{\lambda_j}{2}) \right|; \qquad \hat{\varepsilon}(\lambda_j) = \frac{\partial}{\partial \tau} \log g(\lambda_j; \hat{\tau});$$

$$\lambda_j = \frac{2\pi j}{T}; \qquad \hat{\tau} = \arg\min_{\tau \in T^*} \sigma^2(\tau),$$

where T^* is a compact subset of the R^q Euclidean space. $I(\lambda_j)$ is the periodogram of u_t evaluated under the null hypothesis, i.e.,

$$\hat{u}_t = (1 - L)^{d_o} y_t - \hat{\gamma}' w_t; \qquad \hat{\gamma} = (\hat{\alpha}, \hat{\beta})';$$

$$\hat{\gamma} = \left(\sum_{t=1}^{T} w_t w_t' \right)^{-1} \sum_{t=1}^{T} w_t (1 - L)^{d_o} y_t;$$

$$w_t = (1 - L)^{d_o} z_t, \qquad z_t = (1, t)',$$

and the function g appearing in the equation above is a known function obtained from the spectral density function of u_t,

$$f(\lambda; \sigma^2; \tau) = \frac{\sigma^2}{2\pi} g(\lambda; \tau), \qquad -\pi < \lambda \leq \pi.$$

It has been shown by Robinson (1994) that

$$\hat{r} \to_d N(0,1) \quad \text{as} \quad T \to \infty,$$

and also, that the test is the most efficient one (in the Pitman sense) against local departures from the null hypothesis. We take d_o values equal to –2, –1.99, …, 1.99, 2, and consider the cases of $\alpha = \beta = 0$ a priori (i.e. with no deterministic terms in the undifferenced regression Eqn 8); α unknown and $\beta = 0$ a priori (i.e. with an intercept); and α and β unknown (with a linear time trend).

The results displayed in Table 1 refer to the confidence intervals of those values of d_o where H_0 (Eqn 7) cannot be rejected at the 5 per cent significance level for the three cases of no regressors, an intercept, and an intercept with a linear time trend, assuming that the disturbances u_t in Eqn 9 are white noise, and AR(1) and AR(2) processes. We also display in the table (in parenthesis) the value of d_o producing the lowest statistic in absolute value across d, which should be an approximation to the maximum likelihood estimate. Note that this method is based on the Whittle function, which is an approximation to the likelihood function (Whittle 1954).

Starting with the case of no regressors, the first thing we observe is that the null hypothesis of $I(0)$ stationary x_t (i.e. $d = 0$) is rejected in favour of long memory (d > 0) for the three assumed types of disturbances. The lowest statistic takes place at $d_o = 0.51$ with white noise u_t; at $d_o = 0.38$ with AR(1) u_t, and at $d_o = 0.53$ with AR(2) u_t. Including an intercept and/or a linear time trend, the results are fairly similar in the two cases, and the lowest statistics occur now at values of d

Table 1. Confidence intervals for the values of *d* based on Robinson (1994) in the Australian-averaged annual rainfall data.

u_t / Regressor	No regressors	With an intercept	With a linear trend
White noise	[0.33 **(0.51)** 0.68]	[0.22 **(0.38)** 0.49]	[0.05 **(0.25)** 0.42]
AR(1)	[0.19 **(0.38)** 0.69]	[0.09 **(0.25)** 0.45]	[0.06 **(0.18)** 0.36]
AR(2)	[0.36 **(0.53)** 0.78]	[0.29 **(0.50)** 0.81]	[0.21 **(0.41)** 0.78]

In bold, the values of *d* producing the lowest statistics

Table 2. Selected model for each of the specifications in Table 1.

u_t / Regressor	No regressors	With an intercept	With a linear trend
White noise	$(1 - L)^{0.51} y_t = \varepsilon_t$	$y_t = 99.738 + x_t$ (0.010) $(1 - L)^{0.38} x_t = \varepsilon_t$	$y_t = 139.579 - 4.128t + x_t$ (0.032) (0.000) $(1 - L)^{0.25} x_t = \varepsilon_t$
AR(1)	$(1 - L)^{0.38} y_t = u_t$ $u_t = 0.126 u_{t-1} + \varepsilon_t$	$y_t = 33.230 + x_t$ (0.158) $(1 - L)^{0.25} x_t = u_t$ $u_t = 0.161 u_{t-1} + \varepsilon_t$	$y_t = 87.329 - 1.421t + x_t$ (0.002) (0.001) $(1 - L)^{0.18} x_t = u_t$ $u_t = 0.161 u_{t-1} + \varepsilon_t$
AR(2)	$(1 - L)^{0.73} y_t = u_t$ $u_t = -0.126 u_{t-1} -$ $0.294 u_{t-2} + \varepsilon_t$	$y_t = 191.45 + x_t$ (0.048) $(1 - L)^{0.50} x_t = u_t$ $u_t = -0.088 u_{t-1} -$ $0.265 u_{t-2} + \varepsilon_t$	$y_t = 250.98 - 3.530t + x_t$ (0.006) (0.021) $(1 - L)^{0.20} x_t = u_t$ $u_t = -0.069 u_{t-1} -$ $0.267 u_{t-2} + \varepsilon_t$

p-values in parentheses. The regressor *t* in the last column is a time variable, with *t*=1 for the first observation and *t*=T for the final one.

slightly smaller than in the previous case. In any case, the results presented so far seem to indicate that the national rainfall series is fractionally integrated or *I(d)*, with the estimated value of *d* constrained between zero and one.

In Table 2 we display for each type of deterministic term (no regressors, an intercept, and an intercept with a linear trend), and each type of disturbance (white noise, AR(1) and AR(2)) the selected models associated to the lowest statistics (as shown in Table 1). The most striking feature observed in this table is that the coefficient associated with the time trend is negative and statistically significant for the three types of disturbances (see column four in Table 2). This negative coefficient may appear surprising in view of the plot in Fig. 1 where a slight increasing trend is observed. However, this is a consequence of the long memory structure obtained for the detrended process x_t. In fact, imposing $I(0)$ x_t, (i.e. $x_t = u_t$), and estimating a linear trend (i.e. no fractional differencing), the resulting models are

$$y_t = -40.695 + 0.753t + u_t$$
$$(0.010)\quad(0.003)$$

(*p*-values in parenthesis) in case of white noise u_t, and

$$y_t = -40.322 + 0.719t + u_t;$$
$$(0.008)\quad(0.003)$$

$$u_t = 0.210 u_{t-1} + \varepsilon_t,$$

with AR(1) disturbances. Note, however that in these two cases along with the AR(2) case, the models are rejected according to the results reported in Table 1.

Table 3 displays the results for the regional series assuming that the disturbances are white noise. We note significant differences across the series. Thus, for example, the lowest degree of persistence is obtained for the Murray-Darling Basin, with values of *d* around zero. On the other hand, northern Australia presents the highest values. For the remaining four regions, the values seem to be very sensitive to the choice of the deterministic terms, obtaining different results whether or not we include no regressors, an intercept or an intercept with a linear trend in the model. Thus, for southwestern Australia, the estimated value of *d* is strictly above zero if no regressors or only an intercept is included in the model, however, imposing a linear time trend, *d* is found to be –0.04. The same happens for the eastern, southern and southeastern regions, with values of *d* around 0.30 with no regressors and close to zero in the remaining cases.

Tables 4 and 5 display the results for the same regional data as in Table 3 but assuming autocorrelated disturbances. In Table 4 we assume that u_t is AR(1). In Table 5 we consider a more general type of autocorrelation structure, based on the model of Bloomfield (1973). This is a non-parametric approach that produces autocorrelations decaying exponentially as in the AR(MA) case. We observe in these two tables that practically all values are in the *I*(0) region, the only ex-

Table 3. Confidence intervals for the values of d based on Robinson (1994) in the regionally-averaged rainfall data, based on white noise disturbances.

Region / Regressor	No regressors	With an intercept	With a linear trend
Eastern Australia	[0.23 **(0.38)** 0.56]	[0.00 **(0.08)** 0.20]	[–0.01 **(0.07)** 0.28]
Murray-Darling Basin	[–0.01 **(0.02)** 0.13]	[–0.01 **(0.02)** 0.12]	[–0.05 **(0.00)** 0.16]
Northern Australia	[0.31 **(0.49)** 0.67]	[0.24 **(0.38)** 0.51]	[0.09 **(0.28)** 0.43]
Southeastern Australia	[0.24 **(0.31)** 0.49]	[0.00 **(0.02)** 0.08]	[–0.03 **(0.01)** 0.16]
Southern Australia	[0.25 **(0.34)** 0.53]	[–0.01 **(0.02)** 0.21]	[–0.07 **(–0.03)** 0.26]
Southwestern Australia	[0.33 **(0.43)** 0.55]	[0.26 **(0.34)** 0.40]	[–0.06 **(–0.04)** 0.01]

In bold, the values of *d* producing the lowest statistics

Table 4. Confidence intervals for the values of d based on Robinson (1994) in the regionally-averaged rainfall data based on AR(1) disturbances.

Region / Regressor	No regressors	With an intercept	With a linear trend
Eastern Australia	[–0.04 **(0.00)** 0.07]	[–0.04 **(0.00)** 0.07]	[–0.09 **(–0.03)** 0.22]
Murray-Darling Basin	[–0.02 **(0.01)** 0.09]	[–0.02 **(0.00)** 0.10]	[–0.07 **(–0.02)** 0.20]
Northern Australia	[0.28 **(0.48)** 0.71]	[0.23 **(0.34)** 0.50]	[0.02 **(0.15)** 0.41]
Southeastern Australia	[–0.03 **(0.01)** 0.08]	[–0.03 **(0.01)** 0.08]	[–0.04 **(0.00)** 0.23]
Southern Australia	[–0.03 **(0.00)** 0.04]	[–0.03 **(0.00)** 0.04]	[–0.10 **(–0.06)** 0.01]
Southwestern Australia	[–0.04 **(0.00)** 0.05]	[–0.04 **(0.01)** 0.05]	[–0.05 **(–0.01)** 0.04]

In bold, the values of *d* producing the lowest statistics

Table 5: Confidence intervals for the values of d based on Robinson (1994) in the regionally-averaged rainfall data based on Bloomfield (1) disturbances.

Region / Regressor	No regressors	With an intercept	With a linear trend
Eastern Australia	[–0.04 **(0.00)** 0.08]	[–0.19 **(–0.01)** 0.26]	[–0.43 **(–0.12)** 0.24]
Murray-Darling Basin	[–0.03 **(0.01)** 0.10]	[–0.12 **(0.05)** 0.12]	[–0.42 **(–0.11)** 0.25]
Northern Australia	[0.22 **(0.34)** 0.73]	[–0.12 **(0.08)** 0.31]	[–0.34 **(0.07)** 0.37]
Southeastern Australia	[–0.03 **(0.02)** 0.13]	[–0.08 **(0.10)** 0.34]	[–0.24 **(0.01)** 0.34]
Southern Australia	[–0.04 **(0.01)** 0.04]	[–0.17 **(–0.03)** 0.16]	[–0.07 **(–0.19)** 0.01]
Southwestern Australia	[–0.05 **(0.01)** 0.07]	[–0.01 **(0.12)** 0.31]	[–0.26 **(–0.11)** 0.16]

In bold, the values of *d* producing the lowest statistics

Table 6. Estimation based on fractional integration with a linear trend and a single break (Australian-averaged annual rainfall data)

u_t	T_b	First subsample				Second subsample			
		d_1	α_1	β_1	ρ_1	d_2	α_2	β_2	ρ_2
Wh. Noise	1973	0.00	**432.64 (0.000)**	0.085 (0.826)	—	0.37	**876.90 (0.001)**	–3.887 (0.190)	—
AR (1)	1973	0.00	**430.00 (0.000)**	0.256 (0.511)	–0.016	0.00	**386.18 (0.039)**	1.053 (0.757)	0.310

p-values in parenthesis. In bold, significant coefficients at the 5% level.

Table 7. Estimation based on fractional integration with an intercept and a single break (Australian-averaged annual rainfall data)

u_t	T_b	First subsample			Second subsample		
		d_1	α_1	ρ_1	d_2	α_2	ρ_2
Wh. Noise	1973	0.00	**435.81 (0.000)**	—	0.41	**548.73 (0.000)**	—
AR (1)	1973	0.00	**439.75 (0.000)**	−0.016	0.00	**486.08 (0.000)**	0.309

p-values in parenthesis. In bold, significant coefficients at the 5% level.

Table 8: Estimation based on fractional integration with an intercept and a single break and white noise residuals (Regionally-averaged annual rainfall data)

Region	T_b	First subsample		Second subsample	
		d_1	α_1	d_2	α_2
Eastern Australia	1973	0.00	587.32 (0.000)	0.45	709.35 (0.000)
Murray-Darling Basin	1973	0.00	464.50 (0.000)	0.07	501.75 (0.000)
Northern Australia	1973	0.00	490.35 (0.000)	0.37	627.33 (0.000)
Southeastern Australia	1968	0.00	588.58 (0.000)	0.27	620.00 (0.000)
Southern Australia	1973	0.00	370.75 (0.000)	0.13	408.77 (0.000)
Southwestern Australia	1969	0.00	681.38 (0.000)	0.00	616.20 (0.000)

p-values in parenthesis.

Table 9: Estimation based on fractional integration with an intercept and a single break and AR(1) residuals (Regionally-averaged annual rainfall data)

Region	T_b	First subsample			Second subsample		
		d_1	α_1	ρ_1	d_2	α_2	ρ_2
Eastern Australia	1973	0.00	594.19 (0.000)	0.037	0.00	621.88 (0.000)	0.369
Murray-Darling Basin	1973	0.00	468.22 (0.000)	0.019	0.00	494.11 (0.000)	0.069
Northern Australia	1973	0.00	493.85 (0.000)	0.042	0.00	570.07 (0.000)	0.340
Southeastern Australia	1968	0.04	593.31 (0.000)	0.047	0.00	611.48 (0.000)	0.012
Southern Australia	1973	0.00	372.19 (0.000)	0.024	0.00	399.46 (0.000)	0.110
Southwestern Australia	1969	0.00	679.62 (0.000)	−0.199	0.00	621.72 (0.000)	−0.229

p-values in parenthesis.

ception being northern Australia, where positive orders of integration are obtained in practically all cases.

The results presented in Tables 3 to 5 indicate substantial differences depending on the inclusion or not of intercept and/or linear trends. Thus, in order to check if the time trends are really required, it may be of interest to consider a joint test of

$$H_0: \ d = d_o \ \text{and} \ \beta = 0, \qquad \qquad ...(11)$$

in the model given by Eqns 8 and 9. This possibility is not addressed in Robinson (1994), though Gil-Alana and Robinson (1997) derived an LM test of Eqn 11 against the alternatives

$$H_a: \ d \ne d_o \ \text{or} \ \beta \ne 0, \qquad \qquad ...(12)$$

as follows. We consider the regression model

$$y_t = \beta' z_t + x_t, \ t = 1, 2, ...$$

with the vector partitions $z_t = (z_{At}{}^T, z_{Bt}{}^T)T$, $\beta = (\beta_A{}^T, \beta_B{}^T)^T$, and we want to test $H_0: d = d_o$ and $\beta_B = \beta_{Bo}$. Then, an LM statistic may be shown to be $\hat{r}^2$ (see Eqn 10) plus

$$\sum_{t=1}^{T} \hat{u}_t w_{Bt}{}^T \left(\sum_{t=1}^{T} w_{Bt} w_{Bt}{}^T - \sum_{t=1}^{T} w_{Bt} w_{At}{}^T \left(\sum_{t=1}^{T} w_{At} w_{At}{}^T \right)^{-1} \sum_{t=1}^{T} w_{At} w_{Bt}{}^T \right)^{-1} \sum_{t=1}^{T} \hat{u}_t w_{Bt}$$

$$...(13)$$

with $w_t = (w_{At}{}^T, w_{Bt}{}^T)^T = w_t = (1-L)^{do} z_t$,

$$\hat{u}_t = (1-L)^{do} y_t - (\widetilde{\beta}_A{}^T, \beta_{Bo}{}^T) w_t; \ \widetilde{\beta}_A = \left(\sum_{t=1}^{T} w_{At} w_{At}{}^T \right)^{-1} \sum_{t=1}^{T} w_{At} (1-L)^{do} y_t,$$

$$\hat{\sigma}^2 = T^{-1} \sum_{t=1}^{T} \hat{u}_t{}^2,$$

and $\hat{r}^2$ is calculated as above but using the $\hat{u}_t$ just defined. If the dimension of z_{Bt} is q_B, then we compare Eqn 13 with the upper tail of the $\chi^2_{1+q_B}$ distribution. In our case, testing Eqn 11 against Eqn 12 in Eqns 8 and 9 with $z_t = (1, t)^T$, we have $q_B = 1$, $z_{At} = 1$, $z_{Bt} = t$ for $t \ge 1$. Although we do not report the results in the paper, the time trends were found to be not required in all cases.

Next, we allow for the existence of a changing trend in the regression model Eqn 8 and perform the procedure described in the statistical model section, first including a linear time trend, and then only with an intercept for each subsample.

Starting with the Australian-averaged data, the results with a linear time trend are displayed in Table 6. The break-date is found to be in 1973 for the two cases of white noise and AR(1) disturbances. If u_t is white noise, the orders of integration are zero and 0.37 respectively before and after the break, implying a stronger degree of association between the observations after 1973. The coefficient for the time trend is positive before the break and negative after the break though statistically insignificant in the two cases. If we permit weak autocorrelation in the disturbance term the two orders of integration are equal to zero, and the AR correlation coefficients are respectively –0.016 and 0.310. The time trend coefficients are now both positive though once more statistically insignificant.

Given the lack of significance of the time trend coefficients in the results presented in Table 6, for Table 7 we assume that the model is fully described with an intercept and the fractionally integrated processes. Once more the disturbances are described in terms of white noise and AR(1) processes.[3] The estimated dates for the breaks are obtained at exactly the same date as in the previous cases, i.e. 1973. If u_t is white noise, $d_1 = 0$ and $d_2 = 0.41$, however, with autocorrelated disturbances, both orders of integration are equal to zero, with AR coefficients, $\rho_1 = –0.016$ and $\rho_2 = 0.309$. Thus, it is concluded that the level of association between the observations is higher once the break in the early 1970s is taken into account. This level of correlation can be described throughout two competing models: the fractional differencing and the autoregression. Thus, if we do not permit autoregressions in the disturbance term, the estimated d after the break is 0.41, however if AR is allowed, the long memory property disappears, d becomes exactly zero and all the association between the observations is then described through the AR coefficient. We conducted Likelihood Ratio (LR) tests on both subsamples, and the evidence supports the AR(1) specifications in the two subsamples. In this case the two subsamples are $I(0)$ and we observe an increase in the value of the intercept (representing now the mean of the process) after the break.

We finally employ the procedure for changing intercepts with $I(d)$ on the regionally-averaged data. We first report the results for the case of white noise disturbances (Table 8). The break-date takes place at 1973 in four regions (Murray-Darling Basin, eastern, northern and southern Australia); it occurs at 1968 for southeastern Australia and at 1969 for southwestern Australia. As with the global data, the orders of integration are exactly zero for the first subsamples, and oscillate widely across regions during the second part of the samples. Thus, it is zero for southwestern Australia; 0.07 for the Murray-Darling Basin; 0.13 for southern Australia; 0.27 for southeastern Australia, 0.37 for northern Australia and 0.45 for eastern Australia. Thus, we observe substantial differences in terms of the degree of persistence across the regions during the second part of the samples.

Allowing for autocorrelated (AR(1)) disturbances, the results are displayed in Table 9. The first thing we observe is that the break-dates take place at exactly the same dates as in the previous case of white noise u_t, and the orders of integration are found to be zero in all except one single case

---

[3] Including higher AR orders for the error term produced generally the same results.

corresponding to southeastern Australia during the first subsample. In general, the magnitudes of the AR coefficients are higher during the second subsamples implying a higher degree of persistence after the break. Also, the intercepts are higher in the second subsamples with the exception of southwestern Australia where the mean value is higher in the first subsample.

The break-date results found in these national and regional data may differ in some cases from those obtained in previous works where the break-date was found in more recent periods (around 1997) (e.g. Chambers 2003; Nicholls 2003). However, since the frequency of the data employed in this work is annual, the inclusion of a break in the late 1990s would imply the use of very few observations in the second subsamples, invalidating the analysis based on fractional (or even non-fractional) integration. It is in fact a standard practice in the time series literature to remove the first and last 10% of observations to detect structural breaks (e.g. Bai and Perron 1998; Banerjee and Urga 2005).

Finally, it should be noted that the methodology described in the statistical model section can be extended to the case of more than a single break. However, for the validity of the type of long-memory (fractional integration) model we use in this application, it is necessary that the data span a sufficiently long period of time to detect the dependence across time of the observations; given the sample size of the Australian rainfall series employed here, the inclusion of two or more breaks would result in relatively short subsamples, thereby invalidating the analysis based on fractional integration.

Concluding comments

In this article we have examined the annual Australian-averaged rainfall data for the time period 1900-2006 using a new statistical approach based on fractional integration with a structural break. We use a procedure developed by Gil-Alana (2008) that permits us to estimate deterministic terms and fractional orders of integration for each subsample, with the time of the break being endogenously determined by the model.

The results indicate that if no break is taken into account the series displays long memory and the order of integration ranges between 0.18 and 0.51 depending on the inclusion or not of deterministic terms and the type of $I(0)$ disturbances employed in the model. If a single break is imposed, it takes place in the early 1970s. The first subsample is found to be $I(0)$ with little or no autocorrelation, while the second one is more persistent (or autocorrelated), which may be modelled either with a fractional process or through a simple AR(1) process. Looking at the regional data we note substantial differences in the degree of persistence across the regions, especially after the breaks.

We can conclude the analysis in this paper by saying that the level of persistence in the annual Australian-averaged rainfall data since 1900 has increased during the last thirty years. The results seem to suggest the existence of two competing models, one based on purely fractional $I(d)$ models, and the simple AR(1) specification. In both models, the degree of dependence is described by a single parameter, d in the fractional case, and ρ in the AR model. For the national data if no autocorrelation is permitted, d is much higher after the break. If AR(1) models are considered, the AR coefficients are also higher during the second subsamples. The same happens with respect to the regional data, and we see that even in the non-fractional case with autocorrelated disturbances the AR coefficients are higher in the second subsamples in all except one case (southeastern Australia).

Other authors have also examined Australian rainfall data from a time series viewpoint. Thus, for example, Whiting et al. (2003) examined Australian rainfall data in four capital cities: Sydney (1859 – 1999); Melbourne (1856 – 1999); Brisbane (1860 – 1996) and Perth (1876 – 1991), and first conducted regressions on time under the assumption that the error term was a pure noise process. The time trend coefficients were found to be statistically insignificant in the four cases. Using simple long range dependence techniques it was found that Sydney rainfall data exhibit long memory behaviour. They also employed a two-state hidden Markov model, assuming that climate characteristics fluctuate between "wet" and "dry" periods. In another paper, Simmons and Hope (1997) examined the degree of persistence in four Australian regions (northeast; southeast; northwest and southwest) by simply performing autocorrelations on the anomalous time series. They found significant memory, especially in the southern regions, and lack of persistence in the northwest. These authors, however, do not present a model to describe such dependence. With respect to the break-date, some authors argue that in some regions it should be earlier than reported here. Thus, for example, Khan (2007) established a break around 1950 for southeastern Australia, though with a possible break in the late 70s (consistent with our results) for western and southern Australia. A final interesting issue is the stability of our results for the time period assessed. For this purpose, we also conducted the break tests in different truncated versions of the series. In particular, we examined the time periods 1905- 2006; 1910 – 2006; 1905 – 2000 and 1910 – 2000. For the first two sub samples, the break took place at exactly the same date as the one reported in the paper. However, for the other two sub samples, the break occurred earlier (in the late 1960s and early 1970s), which may be related to the shorter number of observations after the breaks.

Acknowledgments

References

Bai, J. and Perron, P. 1998. Estimating and testing linear models with multiple structural changes. *Econometrica*, 66, 47-78.

Banerjee, A. and Urga, G. 2005. Modelling structural break, long memory and stock market volatility. *J. Econometrics, 129,* 1-34.

Beran, J. 1994. *Statistics for Long-Memory Processes.* London: Chapman and Hall.

Bloomfield, P. 1973. An exponential model in the spectrum of a scalar time series. *Biometrika, 60,* 217-26.

Chambers, L.E. 2003. Southern Australian rainfall variability and trends, Bureau of Meteorology Research Centre, *BMRC Research Report No. 92.*

Chiew, F.H.S. and McMahon, T.A. 2003. El Niño Southern Oscillation and the Australian rainfall and streamflow. *Aust. J. Water Resources, 6,* 115-29.

Dueker, M.J. and Asea, P.K. 1998. *Non-monotonic long memory dynamics in black market premia.* Typescript, The Federal Reserve Bank of St. Louis.

Gil-Alana, L.A. and Robinson, P.M. 1997. Testing of unit roots and other nonstationary hypotheses in macroeconomic time series. *J. Econometrics, 80,* 241-68.

Gil-Alana, L.A. 2008. Fractional integration and structural breaks at unknown periods of time. *J. Time Series Analysis, 29,* 163-85.

Granger, C.W.J. and Ding, Z. 1996. Varieties of long memory models. *J. Econometrics, 73,* 61-77.

Hurst, H.E. 1951. Long-term storage capacity of reservoirs. *Trans. Amer. Soc. Civ. Eng., 116,* 770-99.

Khan, S. 2007. *Climate change or shift? How it affects the Australian community,* International Centre of Water for Food Security, Charles Sturt University / UNESCO IHP-HELP.

Lavery, B., Kariko, A. and Nicholls, N. 1992. A historical rainfall data set for Australia. *Aust. Met. Mag., 40,* 33-9.

Lavery, B., Joung, G. and Nicholls, N. 1997. An extended high-quality historical rainfall dataset for Australia. *Aust. Met. Mag., 46,* 27-38.

Mandelbrot, B. 1977. *Fractals, forms, chance and dimensions.* New York, Free Press.

McBride, J.L. and Nicholls, N. 1983. Seasonal relationships between Australian rainfall and the southern oscillation. *Mon. Weath. Rev., 111,* 1998-2004.

Montanari, A. and Rosso, R. 1997. Fractionally differenced ARIMA models applied to hydrologic time series. Identification, estimation and simulation. *Water Resources Research, 33,* 1035-44.

Montanari, A., Rosso R. and Taqqu, M. 1996. Some long-run properties of rainfall records in Italy. *J. Geophys. Res., 101,* 431-8.

Nicholls, N. 2003. Continued anomalous warming in Australia. *Geophys. Res. Lett., 30,* 1370.

Robinson, P.M. 1994. Efficient tests of nonstationary hypotheses. *J. Amer. Statist. Assoc., 89,* 1420-37.

Simmonds, I. and Hope, P. 1997. Persistence characteristics of Australian rainfall anomalies. *Int. J. Climatol., 17,* 597-613.

Srikanthan, R., Peel, M.C., Pegram, G.G.S. and McMahon, T.A. 2007. Low frequency climate variability and stochastic modelling of annual rainfall data. *Geophys. Res. Abs., 9.*

Whiting, J.P., Lambert, M.F. and Metcalfe, A.V. 2003. Modelling persistence in annual Australian point rainfall. *Hydrol. Earth Sys. Sci., 7,* 197-211.

Whittle, P. 1954. On stationary processes in the plane. *Biometrika, 41,* 434-49.