Vol. 21: 127–140, 2002 | CLIMATE RESEARCH Clim Res | Published June 14

Long memory in surface air temperature: detection, modeling, and application to weather derivative valuation

Rodrigo Caballero[1,*], Stephen Jewson[2], Anders Brix[2]

[1]Danish Center for Earth System Science, University of Copenhagen, Juliane Maries Vej 30, 2100 Copenhagen Ø, Denmark
[2]Risk Management Solutions, London, United Kingdom

ABSTRACT: Three multidecadal daily time series of mid-latitude near-surface air temperature are analysed. Long-range dependence can be detected in all 3 time series with 95% statistical significance. It is shown that fractionally integrated time-series models can accurately and parsimoniously reproduce the autocovariance structure of the observed data. The concept of weather derivatives is introduced and problems surrounding their pricing are discussed. It is shown that the fractionally integrated time-series models provide much more accurate pricing as compared with traditional autoregressive models employing a similar number of parameters. Finally, it is suggested that a simple explanation for the presence of long memory in the time series may be given in terms of aggregation of several short-memory processes.

KEY WORDS: Surface temperature · Long memory · Weather derivative

1. INTRODUCTION

The daily-mean near-surface air temperature (SAT) measured at a given point fluctuates randomly from day to day. In the mid-latitudes, this is largely due to the ceaseless passage of high- and low-pressure systems, whose associated horizontal and vertical winds bring and remove heat, moisture and clouds to and from the measurement point. These traveling perturbations generally exhibit a characteristic life cycle featuring rapid initial growth fueled by baroclinic energy conversion followed by slower barotropic decay (Simmons & Hoskins 1978). On this basis, one might heuristically model SAT as a simple Ornstein-Uhlenbeck process:

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = \gamma x(t) + \sigma\eta(t) \quad (1)$$

where x is the SAT anomaly, η is Gaussian white noise with unit variance and γ and σ are constants. The white-noise term would represent the rapid excitation of the perturbations, while the first term on the right-hand side assumes that the perturbations are linearly damped with a time constant, γ^{-1}. A reasonable value for this time scale may be obtained by noting that the average eddy kinetic energy of the atmosphere is around 106 J m^{-2}, while the mean potential-to-kinetic energy conversion rate, which must match the dissipation rate, is about 2 W m^{-2} (Peixoto & Ort 1992); thus $\gamma^{-1} \approx 5 \times 10^5$ s $\approx$ 6 d. For processes in discrete time, Eq. (1) may be discretised to give a first-order autoregressive or AR(1) process:

$$x_i = \alpha x_{i-1} + \varepsilon_i \quad (2)$$

with $\alpha = 1 - \gamma\Delta t$, where Δt is the discretisation interval and ε_i is a Gaussian-white-noise process with variance $\sigma^2\Delta t$.

The power spectral density (S) of the process described by Eq. (1) takes the form:

*E-mail: rca@dcess.ku.dk

$$S(\omega) = \frac{\sigma^2}{2\pi}\frac{1}{\omega^2+\gamma^2} \quad (3)$$

where ω is the angular frequency. We thus expect the SAT spectrum to be almost white (constant) at low frequencies and red (negatively sloping) with slope –2 on a log-log plot at high frequencies. The crossover between these 2 regimes should be at $\omega \approx \gamma$, that is, around $\omega = \frac{1}{6}$ rad d^{-1}. How successful is this prediction?

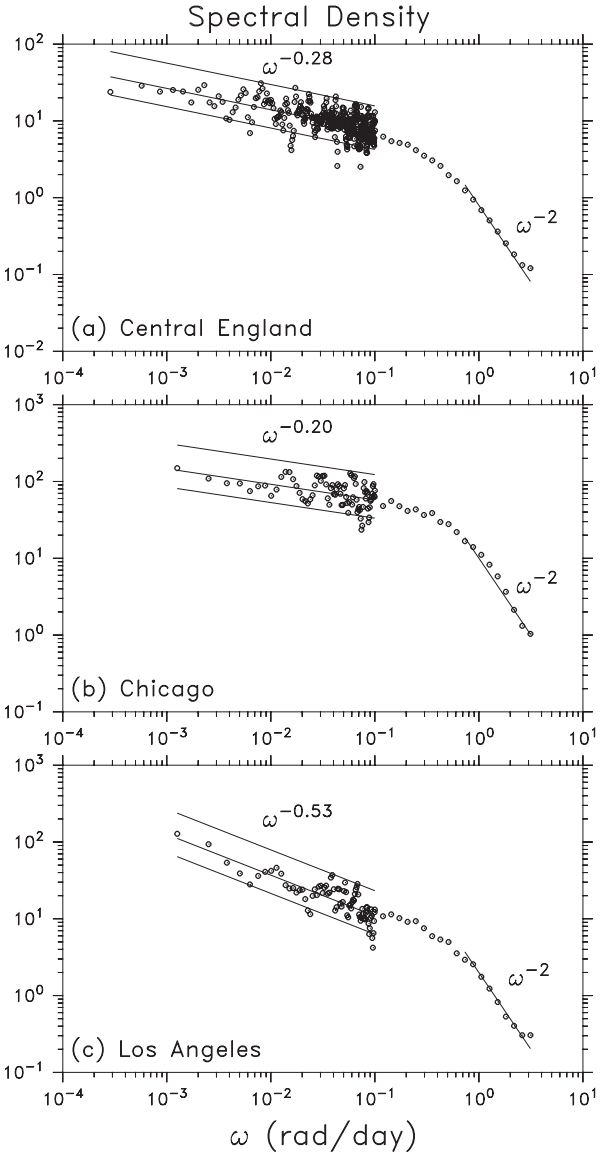

Fig. 1. Power spectral density estimates for SAT measured in (a) Central England, (b) Chicago and (c) Los Angeles. The spectra are smoothed using a Tukey window giving 19 degrees of freedom in all cases. At frequencies above 10^{-1} rad d^{-1}, the spectra are further smoothed by averaging within 20 logarithmically equispaced bins. The data were detrended and deseasonalised in the mean before computing the spectral estimator. The straight lines at low frequencies show a least-squares fit (with the slope indicated on each panel) and an approximate 95% confidence envelope. The line at high frequencies, with slope –2, is shown for comparison

A glance at some observed spectra (Fig. 1; see Section 2 for a description of the data) shows partial success. At high frequencies, the spectra are indeed well approximated by a straight line of slope –2. Further, all spectra show a shoulder or crossover region centered around 6 d. However, the low-frequency region is not white, as expected, but weakly red. There are no statistically significant peaks, and at frequencies below 10^{-1} rad d^{-1} the spectra are well fitted by a straight line with a negative slope.

Processes showing the kind of power law low-frequency behaviour suggested in Fig. 1 are collectively referred to as having 'long memory'. More precisely, a stationary stochastic process is said to exhibit long memory or long-range dependence if there exists a real number, $d \in (0, \frac{1}{2})$, known as the 'intensity' of the long memory, such that:

$$S(\omega) \approx \omega^{-2d} \quad \text{as} \quad \omega > 0 \quad (4)$$

(Beran 1994). If Eq. (4) holds, then the autocorrelation function ρ(τ) will behave as:

$$\rho(\tau) \approx \tau^{2d-1} \quad \text{as} \quad \tau \to \infty$$

where τ is the lag; that is, it will decay as a power law, meaning that highly persistent serial correlations will be present in the data (hence the term 'long memory'). By contrast, when $d = 0$ the autocorrelation decays exponentially and the process is said to have short memory; this is the case for the Ornstein-Uhlenbeck process above. Long memory was first empirically detected by Hurst (1951), who studied Nile river-level data. Since then it has been detected in a host of environmental data series, with examples from geophysics (Mandelbrot & Wallace 1969), hydrology (Montanari et al. 1996), surface winds (Haslett & Raftery 1989), SAT (Bloomfield 1992, Koscielny-Bunde et al. 1996, Pelletier 1997, Syroka & Toumi 2001), mid-tropospheric geopotential heights (Tsonis et al. 1999) and the North Atlantic Oscillation (Stephenson et al. 2000).

A characteristic trait of long-memory processes is that the variance of an N-member sample mean decreases more slowly than N^{-1} (Beran 1989). This can have important consequences for applications. For instance, incorrectly assuming short memory can lead to exaggeratedly narrow confidence intervals for the mean. That was the problem tackled by Haslett & Raftery (1989) in their study of the mean power obtainable from a wind turbine.

In the present paper, we address a similar problem but in the context of a different application, namely weather derivatives. These, as discussed in greater

detail in Section 5, are essentially a form of insurance against fluctuations in weather, of interest to companies whose business is affected by such fluctuations.[1] In order to set the insurance premium (i.e. to value the weather derivative), it is necessary to know the probability distribution of a weather index, typically a seasonal-mean temperature measured at a specified station. One can of course simply look at historical data and obtain suitable estimates. For a number of reasons, however, it is desirable to have a stochastic time-series model which permits the use of Monte Carlo methods.

The classical approach to time-series modeling involves fitting a model of the Box-Jenkins type (see Section 4) using as few parameters as possible. It is well known, however, that these models are subject to the 'overdispersion' problem, i.e. will underestimate the variance of seasonal means (Shea & Madden 1990, Katz & Parlange 1998). Here, we show that the problem may be overcome (at least for our particular application) by employing a generalisation of Box-Jenkins models, known as fractional ARIMA or ARFIMA (Granger & Joyeux 1980, Hosking 1981), which incorporate long memory in a natural way, using the single parameter d. These models offer particular flexibility in that they can capture both high-frequency short-memory behaviour and the long-memory tail using a minimum of parameters.

Our aims in the present study are 3-fold. Firstly, we wish to show that long memory can indeed be detected in SAT time series. We have already presented some evidence for this in Fig. 1, and more is given in Section 3. Secondly, we show that ARFIMA models can accurately and parsimoniously capture the autocorrelation structure of observed SAT time series; this is done in Section 4. Our third aim is to present an example of the application of time-series modeling to weather derivative pricing and compare the performance of ARMA and ARFIMA models having the same number of parameters (Section 5). We comment also on the fact that the autocorrelation structure contains seasonal dependency and outline its effects on the performance of the time-series models. In Section 6, we offer some suggestions as to the possible physical origin of long memory. A summary and conclusions are given in Section 7.

[1]The term 'derivative' is used in finance to indicate a security whose value is based on that of another security, called the 'underlying': an example of a derivative is an 'option', whose value is based on an underlying stock price. An introduction to financial derivatives may be found in Hull (1998). In the case of weather derivatives, the underlying is a temperature or other weather index (see Section 5)

2. DATA AND PREPROCESSING

This study is based on 2 data sets. The first is the daily Central England temperature time series (Parker et al. 1992). It is representative of a roughly triangular area of the United Kingdom enclosed by Preston, London and Bristol. The time series, beginning in 1772 and ending in 1991 (222 yr, 81 084 d), is one of the longest available instrumental records of daily temperature in the world and is thus particularly suitable for examining the asymptotic properties of interest here. The second data set, prepared by Risk Management Solutions on the basis of data provided by NOAA (US National Oceanic and Atmospheric Administration), consists of daily temperature records for the last 50 yr (18 262 d) at 200 weather stations in the US. We focus on 2 stations, Chicago and Los Angeles, which are representative of stations having respectively the minimum and maximum long-memory intensity within the dataset.

In the following sections, we will apply a suite of tests to these time series designed to detect the presence of long memory. From a physical point of view, detecting long memory can only be considered interesting if it reveals something about the 'internal' workings of the climate system. Thus, before applying any tests, we must try to rid the time series from the signature of processes which are 'external' to the climate system. Processes which induce nonstationarity in the mean are particularly problematic, since these are most likely to lead to spurious detection of long memory. There are several such processes:

- Seasonality, which depends on changes in solar forcing and is therefore not internal to the climate system, will give a periodic signal in both mean and variance. It may be trivially removed by Fourier-transforming the raw SAT time series and estimating the seasonal cycle using the amplitude and phase of the Fourier coefficients corresponding to annual and sub-annual periodicities (the only harmonics that we found to give well-defined peaks in the spectrum). A similar procedure was applied to the squared SAT anomalies to estimate and remove the seasonal cycle in the variance.
- Urbanization, whereby weather stations originally in open country are gradually engulfed by nearby towns or cities and their accompanying heat island (Cotton & Pielke 1995). This leads to a strong upward trend in the time series, on the order of several degrees centigrade over the last 50 yr. It is difficult to model the exact form of the trend, which will vary from station to station. Lacking the information to do otherwise, we make the simplest choice, which is to detrend the time series using a linear least-squares fit.

Table 1. Estimates of d ('intensity' of the long memory) obtained by various methods: least-squares regression on the periodogram (P), aggregated variance (AV), differenced variance (DV) and maximum likelihood estimation of an ARFIMA(1,d,1) model (ML). Where available, 95% confidence intervals are shown in parentheses

	P	AV	DV	ML
Central England	0.16 (±0.04)	0.12	0.15	0.20 (±0.02)
Chicago	0.12 (±0.08)	0.09	0.19	0.13 (±0.04)
Los Angeles	0.29 (±0.08)	0.24	0.31	0.23 (±0.02)

- Shifts in station position or instrumentation can lead to abrupt changes in the mean of the time series; a clear example of this (and of the urbanization trend) is discussed in von Storch & Zwiers (1999, p. 9). Much effort has been devoted to eliminating these inhomogeneities in the US dataset. Where possible, station records have been consulted to establish the date of position or instrumentation changes. Unfortunately such records are often incomplete or unavailable, and in those cases a statistical procedure is adopted which attempts to locate the breaks. We did not apply this procedure to the Central England dataset.

In addition, there are smaller effects due to global warming, volcanism and changes in solar activity and stratospheric ozone which will also give secular trends; we assume (again for want of better knowledge) that their aggregate effect is also linear and is removed together with the urbanization trend.

At this point we could in principle also remove the signatures of such well-known phenomena as the El Niño/Southern Oscillation, the North Atlantic Oscillation and so on. We do not do this for 2 reasons. One is that it is technically difficult to do so in a meaningful way—for instance, one could build a regression model between an El Niño index and the SAT time series, but questions arise as to which index is the most suitable and as to the appropriateness of linear regression. The main reason, however, is that these are modes of variability 'internal' to the climate system, and are thus part and parcel of the 'interesting' signal. Some further remarks on the relation between these low-frequency forms of variability and higher frequencies are offered in Sections 6 and 7.

3. DETECTION OF LONG MEMORY: HEURISTIC METHODS

Long memory was first detected using the rescaled range or R/S statistic (Hurst 1951). Since then, a number of similar methods have been developed to detect long memory in time series (see Taqqu et al. 1995 for a list). Each method produces an estimate of d; if $d > 0$, we say the time series has long memory. The methods are intuitive and easy to apply, but they are 'heuristic' in that they generally do not permit confidence interval estimation (though the periodogram method, see below, is an exception). We consider 3 of these methods here; motivation for these particular choices is given below.

3.1. Periodogram method

In Eq. (4) we defined long memory as an asymptotic property of the spectral density. Thus, a direct way to test for long memory is to compute a spectral estimator and fit a straight line (on a log-log plot) to the low-frequency end; we then have $d = -b/2$, where b is the slope of the line. We select this method because it is exceptional in that a procedure for confidence interval estimation is known (Beran 1994, p. 98), provided the unsmoothed periodogram is used (note that a smoothed periodogram, which is more suitable for display purposes, was used in Fig. 1). We applied the method to our 3 time series using classical least-squares regression in the frequency interval 10^{-4} to 10^{-1} rad d^{-1}. Results are reported in Table 1. We can infer with 95% confidence that all 3 series display long memory. Changing the regression interval changes the numerical values somewhat but does not change this qualitative conclusion.

3.2. Aggregated variance and differenced variance methods

As mentioned in the Section 1, a key property of long-memory processes is that the variance of sample means decreases slowly with sample size. In fact, it can be shown (Beran 1989) that given N data points, X_i, $i = 1, \ldots, N$:

$$\mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) \approx N^{2d-1} \quad \text{as} \quad N \to \infty$$

This suggests the following method for estimating d. Divide the series into N/m blocks of size m and compute the mean for each block:

$$x_k(m) = \frac{1}{m}\sum_{i=(k-1)m+1}^{km} X_i \qquad k = 1,\ldots, N/m$$

and the variance of the block mean:

$$s^2(m) = \frac{1}{N/m - 1}\sum_{k=1}^{N/m}\left[x_k(m) - \bar{x}\right]^2$$

where $\bar{x}$ indicates an overall mean. Now a log-log plot of $s^2(m)$ against m should yield a straight line with slope $2d - 1$. This is known as the aggregated variance method.

A shortcoming of this and other heuristic estimators is that inhomogeneity in the data (see Section 2) can produce a positive value of d even in the absence of long memory. A modification of the above method, known as the differenced variance method, avoids this problem. The idea is to study the first-order difference of the above variances:

$$\Delta s^2(m) = s^2(m + 1) - s^2(m)$$

It can be shown (Teverovsky & Taqqu 1997) that a log-log plot of this quantity against m will again asymptotically produce a straight line with slope $2d - 1$, but this time the value of d will be free from the effects of inhomogeneity in the mean.[2]

We have applied both of the above procedures to our 3 time series using 50 logarithmically equispaced values of m. The results are plotted in Fig. 2. Given the considerable scatter in the points at high m, we use robust fitting in the interval $30 < m < 5000$ d to estimate d. Estimated values are given in Table 1. We note that the values of d estimated with the differenced variance method are all greater than those obtained with the aggregated variance, so that inhomogeneity does not seem to play a role. We note also that these estimates are all consistent (within the 95% confidence interval) with the periodogram estimates. Again, the values obtained here are sensitive to the choice of regression interval, but not so much as to invalidate the above qualitative conclusions.

4. MODELING SAT TIME SERIES

The classical approach to modeling discrete time series is through the use of autoregressive integrated moving average (ARIMA) models, popularised by Box & Jenkins (1970). In their notation, they take the general form:

$$\phi(B)\,(1 - B)^d X_i = \psi(B)\varepsilon_i \qquad (5)$$

Here, ε is a white noise process and B is the backstep operator:

$$BX_i = X_{i-1}$$

while $\phi(B)$ and $\psi(B)$ are polynomials in B:

$$\phi(B) = a_0 + a_1 B + a_2 B^2 + \ldots$$

$$\psi(B) = b_0 + b_1 B + b_2 B^2 + \ldots$$

which represent the autoregressive and moving average parts respectively. The term:

$$(1 - B)^d$$

where d is zero or a positive integer, represents a finite-difference time derivative of order d; the interpretation is that one should first compute the process with $d = 0$, and then sum (integrate) it d times to obtain the full process. ARIMA processes are non-stationary for $d = 1$. Since the processes we are dealing with here are stationary, we need only consider the case $d = 0$. Then, if ϕ and ψ are of degree p and q respectively, the process is denoted ARMA(p,q), or AR(p) if $q = 0$.

The autocorrelation function of an ARMA(p,q) process will decay exponentially to zero for lags greater than max(p,q). For this reason, they are unsuitable for modeling long-memory time series; in order to obtain appreciable correlation at, say, Lag 100, it is necessary to use an ARMA process of order 100, which means fitting 100 parameters. This contravenes the general principle of parsimony (Box & Jenkins 1970), according to which one should always seek a model which will will adequately fit the data with the bare minimum of parameters.

It turns out that daily temperature time series can be accurately and parsimoniously fitted using a more general class of stochastic process known as fractional ARIMA or ARFIMA models. They are defined exactly as in Eq. (5), except that now $0 < d < \frac{1}{2}$; it can be shown (Granger & Joyeux 1980) that in this range the models are stationary and have the long-memory property with intensity d. For $d = \frac{1}{2}$, the models are non-stationary. To make sense of the fractional differencing, the operator is formally expanded as a power series:

$$(1-B)^d = \sum_{k=0}^{\infty} \frac{\Gamma(d+1)}{\Gamma(k+1)\Gamma(d-k+1)} (-1)^k B^k \qquad (6)$$

Thus, an ARFIMA(p,d,q) model is equivalent to an ARMA(8,q) model while using only $p + q + 1$ parameters. The presence of autoregressive and moving-average components allows these models to capture the short-memory high-frequency behaviour, while the slow decay of the coefficients in Eq. (6) controls the long-term behaviour. This allows ARFIMA time series to accurately model long-memory processes while using only a small number of parameters.

[2]There is an interesting parallel between the problem addressed here and those faced in the study of long-range correlations in DNA nucleotides. These were originally studied using the 'fluctuation analysis' method (Peng et al. 1992), which is almost identical to the aggregated variance method. It was later objected that the long-range correlations detected could be trivial due to the well-known 'patchiness' or inhomogeneity of DNA. To get around this, 'detrended fluctuation analysis' was devised (Peng et al. 1994), which is insensitive to the inhomogeneity and is in this sense analogous (though quite different in detail) to the differenced variance method

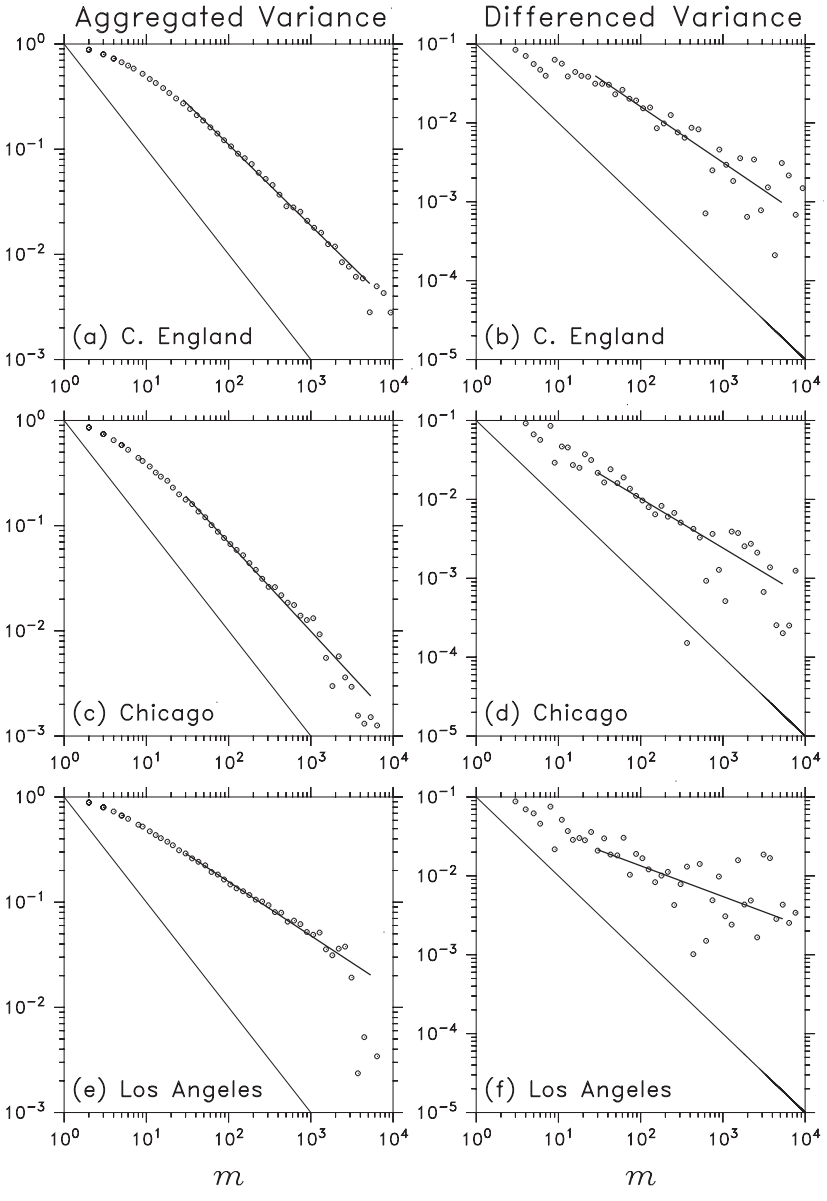

Fig. 2. Aggregated variance and differenced variance for the 3 datasets (detrended and deseasonalised in mean and variance) as a function of the sample size m (see Section 3). The thick solid line shows a robust fit to the data points. The thin line, with slope –1, is shown for comparison

Fitting an ARFIMA model to data is somewhat more involved than with ARMA models. Exact maximum likelihood (ML) estimation is possible but prohibitively expensive from a computational viewpoint for the length of the time series and practical applications we are considering here. Fortunately, a number of efficient approximate ML methods have been developed (see review by Beran 1994). ML methods naturally permit estimation of confidence intervals for the fitted parameters, and hence provide an alternative and more rigorous method for detection of long-range dependence than the heuristic methods discussed in Section 3.

Here we use the approximate ML method proposed by Haslett & Raftery (1989) and available in the 'fracdiff' package of the 'R' statistical computing environment. We used these routines to fit an ARFIMA(1,d,1) model to our 3 data series. This particular ARFIMA model was selected by trial and error. Models with less than 2 ARMA parameters all give a considerably worse fit (as assessed by comparing the model and data autocorrelation functions), while those with more than 2 give no significant improvement in the fit; amongst those with 2 ARMA parameters, ARFIMA(1,d,1) gave the best fit in the 3 cases studied here. Results are shown in Table 1. The results again indicate that long memory is present in all 3 series at the 95% confidence level. We note also that the ML values are compatible with those obtained using the periodogram, within the respective uncertainties.

To show how significant an improvement ARFIMA is over ARMA, we compare the autocorrelation functions of the historical data to those of ARFIMA(1,d,1), AR(3) and AR(20) models fitted to the time series (note that AR models fitted to a given series generally have autocorrelations which decay more slowly than MA or ARMA models with the same number of parameters and are hence better suited to time series with persistent correlations). Results are shown in Fig. 3. In all 3 time series, the ARFIMA model gives a much better fit to the autocorrelation structure than the other models, even at high lags. There is of course some scatter in the observed autocorrelation, but the ARFIMA model appears to give a more-or-less unbiased fit (though with some overestimation in the Central England case). The AR(3) model, on the other hand, quickly drops to zero and substantially underestimates the correlation at lags higher than a few days. The AR(20) model closely follows the observed autocorrelation function up to a lag of 20 d, as expected, but then drops off rapidly, again consistently underestimating persistence at higher lags.

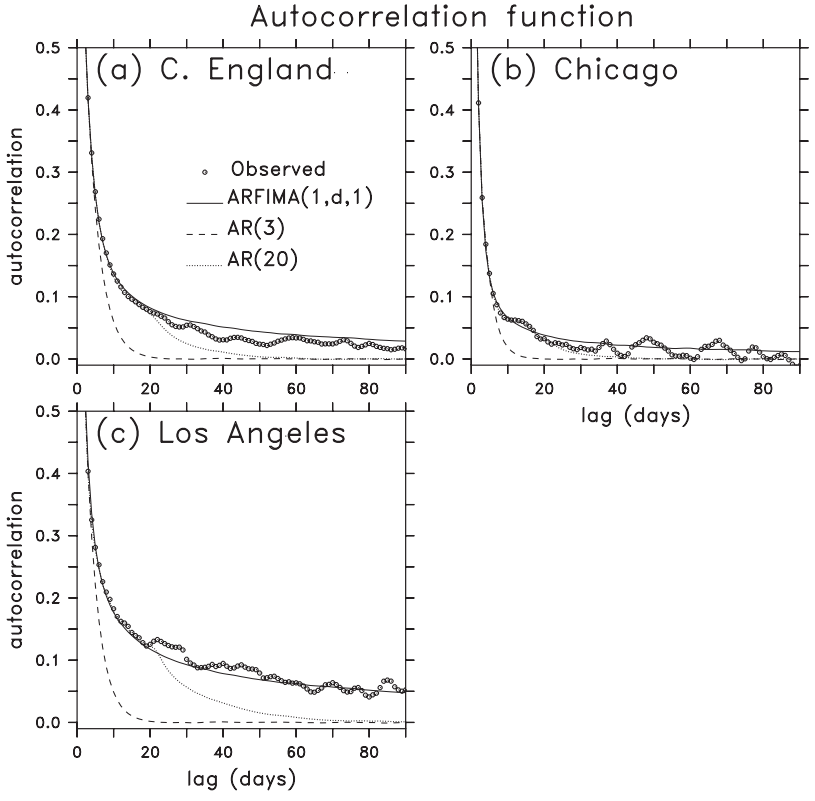

Fig. 3. Autocorrelation functions for the (a) Central England, (b) Chicago and (c) Los Angeles SAT anomaly time series with the observed data and fits for the ARFIMA(1,*d*,1), AR(3) and AR(20) models

5. PRICING WEATHER DERIVATIVES: ARMA VERSUS ARFIMA

5.1. What is a weather derivative?

As noted in the Section 1, weather derivatives are a form of insurance, or in financial terms an instrument which a company can use to hedge its exposure to fluctuations in weather. They are a rather recent development, dating back only to 1997, and are not yet widely known in the meteorological literature (an exception being Zeng 2000).

The basic concept may be understood through a simple example. A company selling gas for heating may want to protect itself against losses due to anomalously mild conditions the following winter. It can then buy an insurance policy (the weather derivative) from a financial institution which agrees to pay out a certain amount if the winter should indeed turn out to be mild. Note that the company thereby reduces its losses in adverse circumstances, but incurs the cost of the insurance premium, thus reducing its gains if the winter turns out to be harsh. The purpose of weather derivatives is to smooth out the temporal fluctuations in the company's revenues. There are a number of financial and commercial reasons why this is beneficial.

Two ingredients are necessary in order to structure a weather derivative contract such as the above: a weather index and a payout function. The weather index, denoted I, characterises the weather over a certain period, called the 'strike' period. It is most commonly based on temperature (though rain- and snowfall and hours of sunshine have also been used). Rather than use raw temperature, it is customary to use heating or cooling degree days (HDDs or CDDs), defined as:

$$\mathrm{HDD}_i = \max(T^\star - T_i, 0) \quad (7)$$

$$\mathrm{CDD}_i = \max(T_i - T^\star, 0) \quad (8)$$

where the subscript i indicates a specific day, T_i is the average temperature measured on that day, and $T^\star$ is a fixed reference temperature (65°F in the US, 18°C in Europe). In the example above, the index might be defined as the total number of HDDs over the winter season:

$$I = \sum_{i=i_f}^{i_f+N-1} \mathrm{HDD}_i$$

where i_f is the first day of the season and N is its length.

The payout function, $Q(I)$, determines how much the financial institution will pay out for a given index outcome. Referring to the example given previously, we expect no payout if the winter is harsh (high I) and an increasing payout the milder the winter (low I). A typical payout structure has the form shown in Fig. 4, featuring a zero-payout threshold and a linear increase in payout below the threshold. The position of the threshold and the slope of the linear part must be agreed on by the 2 parties entering the contract.

Once the definition of the index and the form of the payout function have been stipulated, the actual payout from the contract depends only on the final index value, which itself depends on the weather. What is less clear is how much the financial institution should charge the company, i.e. how to price the weather derivative. The subject of derivative pricing is a complex one and has received much attention over the past decades. For reasons outside the scope of this paper, the usual way to price a weather derivative is to set:

$$S = E[Q] + R \quad (9)$$

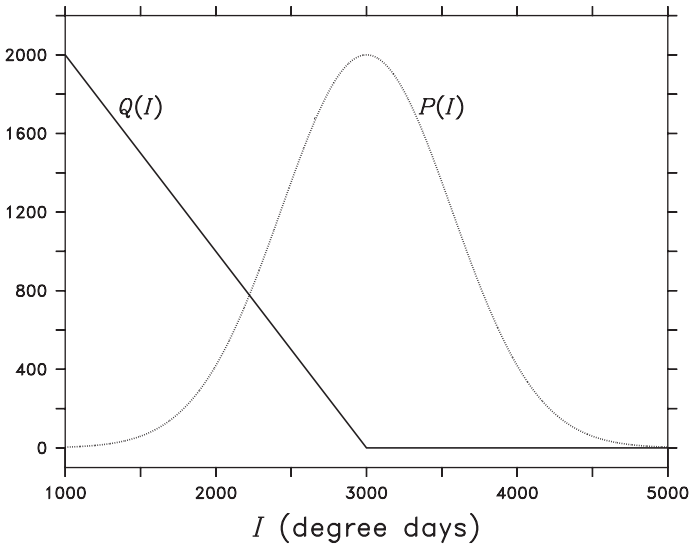

Fig. 4. Plot of a typical payout function $Q(I)$. Overlaid is a hypothetical Gaussian index probability distribution function (pdf), $P(I)$

where S is the price of the derivative. The first term is simply the mean payout:

$$E[Q] = \int_0^\infty Q(I)P(I)\mathrm{d}I$$

$P(I)$ being the probability distribution function (pdf) of the index. The second term, R, is a 'risk premium'—a positive correction which compensates the financial institution for the risk taken in selling the derivative. We will not dwell on its specific form here, but note that it too in general will depend on $P(I)$.

5.2. Why use a time-series model?

We saw above that to price a weather derivative accurately, it is essential to have a good estimate of $P(I)$, or at least of $E[Q]$. There are several possible approaches:

- 'Burn' analysis, which involves evaluating $Q(I)$ for historical values of I and directly computing $E[Q]$.
- Direct modeling of the index distribution, which involves fitting a parametric or non-parametric distribution to the historical values of I.
- Daily modeling (the subject of this paper), which involves fitting a time-series model to the daily data and then using the model to generate a very large sample of synthetic I values.

In current practice, the first 2 methods are by far the most commonly used. Indeed, the daily modeling approach is somewhat more complicated to implement, and there is widespread mistrust of it because of the overdispersion problem. Why then bother with the daily model at all? Firstly, we note that there is no *a priori* reason why a 'good' daily model (i.e. one free from the overdispersion problem) should be any less accurate than the other methods. In fact, in cases where the daily temperature often falls below or above the degree-day threshold, $T^\star$ (see Eqs. 7 & 8), the daily model will make more efficient use of the available data and is likely to be more accurate.

Further, there is at least 1 situation in which daily modeling is clearly the method of choice: the pricing of derivatives within the strike period, a procedure known in financial jargon as 'marking to market'. During the strike period, a new temperature observation becomes available each day, and one can use this information to make a better estimate of the final value of I. One can also make use of deterministic weather forecasts to project information up to 10 d into the future (see, for instance, Jewson 2000). Since SAT time series are highly autocorrelated, it clearly makes sense to use an appropriate time series model to project available information even further into the future, out to the end of the strike period. Accurate marking to market is important because many derivatives are actually most heavily traded during the strike period, and all parties involved in derivatives trading must keep track on a day-by-day basis of their total exposure and risk.

5.3. Factors influencing the price of the derivative

The gross features of $P(I)$, namely its location and width, are captured by its first 2 moments, the mean μ_I and the variance σ_I^2. It is clear from Fig. 4 that, if either the mean is overestimated or the variance underestimated, the derivative will be underpriced. A financial institution trading contracts with such incorrect pricing will, over time, lose money.

Let us then consider what aspects of daily temperature variability influence the values of μ_I and σ_I^2. For the sake of this discussion, we assume that the temperature over the winter period is always below $T^\star$. We can then write:

$$\mathrm{HDD}_i = T^\star - T_i = T^\star - E[T_i] - T'_i$$

where $E[\cdot]$ indicates an expected value and we define the temperature anomalies $T'_i \equiv T_i - E[T_i]$. Then:

$$\mu_I = E[I] = NT^\star - \sum_{i=i_I}^{i_I+N-1} E[T_i] \qquad (10)$$

Thus, to obtain an accurate estimate of the mean index value, μ_I, we need only an accurate estimate of the seasonal cycle, $E[T_i]$.

The situation is more complicated for the variance. We have:

$$\sigma_I^2 = E\left[(I-\mu_I)^2\right] = E\left[\left(\sum_{i=i_f}^{i_f+N-1} T'_i\right)^2\right] = \sigma_T^2 \sum_{i,j=i_f}^{i_f+N-1} \rho_{i,j}$$

where $\sigma_T{}^2$ is the variance and $\rho_{i,j}$ the autocorrelation function of the temperature anomalies. Assuming the process is stationary, the autocorrelation will only depend on the lag k; we then have:

$$\sigma_I^2 = \sigma_T^2\left(N + 2\sum_{k=1}^{N}(N-k)\rho_k\right) \quad (11)$$

Thus, to estimate of the variance of I accurately, we need to estimate not only the variance of the temperature anomalies, but also their autocorrelation structure up to lag N. If the autocorrelation is underestimated, so too will the index variance be. In summary, the minimum requirements for a suitable time-series model are that it should correctly capture the seasonal cycle, the anomaly variance, and the anomaly autocorrelation structure out to lags of a season.

5.4. Comparing the performance of ARMA and ARFIMA

Here we compare the skill of best-fit ARFIMA(1,d,1) and AR(3) models in simulating the index pdf; this is essentially a test of how well they will fare in pricing a derivative based on this index. The comparison is fair in the sense that the models use the same number of parameters. We approach the problem by making the null hypothesis that the model is perfect (that is, that the historical data are actually generated by the model itself) and then try to reject the hypothesis. We will see that the hypothesis is much more easily rejected for AR(3) than for ARFIMA(1,d,1).

In practice, we proceed as follows: The historical data provides us with 50 values of I; we use these to obtain the historical sample mean, $\hat{\mu}_I^H$. We now generate a very large number of index values using the fitted model and compute a close approximation to μ_I^M, the model population mean. Let us set $\Delta = \mu_I^M - \hat{\mu}_I^H$. If Δ is large, then we should reject the model. To establish a rejection threshold, we need an estimate of the sampling fluctuations. This we obtain by dividing up the large model sample into 50-member sub-samples. Each sub-sample provides a sample mean, $\hat{\mu}_I^M$, and we can use these to compute a value, Δ_{99}, such that only 1% of the sub-samples gives $|\mu_I^M - \hat{\mu}_I^M| > \Delta_{99}$. Thus, we can reject the model with 99% confidence if $\Delta > \Delta_{99}$. We can proceed analogously for σ_I^2 and indeed for any parameter we wish.

We restrict the test to the Chicago and Los Angeles stations, which are relevant for real-world derivative pricing. For Chicago, we consider a 5 mo HDD contract spanning November–March; for Los Angeles, we consider a 5 mo CDD contract spanning May–September. We generated 2×10^5 index values to perform the test. Note that the models are fitted to deseasonalised data; before computing synthetic index values, we add the estimated seasonal cycle in mean and variance to the model output.

Results are reported in Table 2. In the case of μ_I, modeled and historical values are in good agreement at both stations for both models. In view of Eq. (10), this essentially means that the seasonal cycle is well estimated. Note that Δ values are almost identical for both models, as they should be. Neither model can be rejected at either of the stations.

Results for σ_I^2 are much more critical. The AR model can be rejected at both stations. Note also that Δ is always negative, which means that the model is underestimating the index variance; this is what we expect from Eq. (11), given the systematic underestimate of the autocorrelation (Fig. 3). The ARFIMA model fares much better at Chicago. It also performs better in Los Angeles (it gives a smaller Δ), but it can be rejected at the 99% confidence level. The reasons for this are discussed below.

Table 2. Mean index value, μ_I, computed from observations and from 2 models. For Chicago (Los Angeles), a 5 mo HDD (CDD) contract covering the period November–March (May–September) is considered. Units everywhere are Fahrenheit degree-days. Δ indicates the difference between the historical and modeled value, as a percentage of the latter. Δ_{99} indicates a 99% confidence level for Δ (that is, only 1% of model-produced 50-member samples will give $|\Delta| > \Delta_{99}$)

	Historical	ARFIMA(1,d,1)			AR(3)		
	$\hat{\mu}_I^H$	μ_I^M	Δ	Δ_{99}	μ_I^M	Δ	Δ_{99}
Chicago	5032	5036	–0.1	2.6	5036	–0.1	2.1
Los Angeles	438	457	4.1	10.3	457	4.0	6.3
	$\hat{\sigma}_I^H$	σ_I^M	Δ	Δ_{99}	σ_I^M	Δ	Δ_{99}
Chicago	359	360	1	24	280	–27	26
Los Angeles	166	126	–32	20	80	–107	25

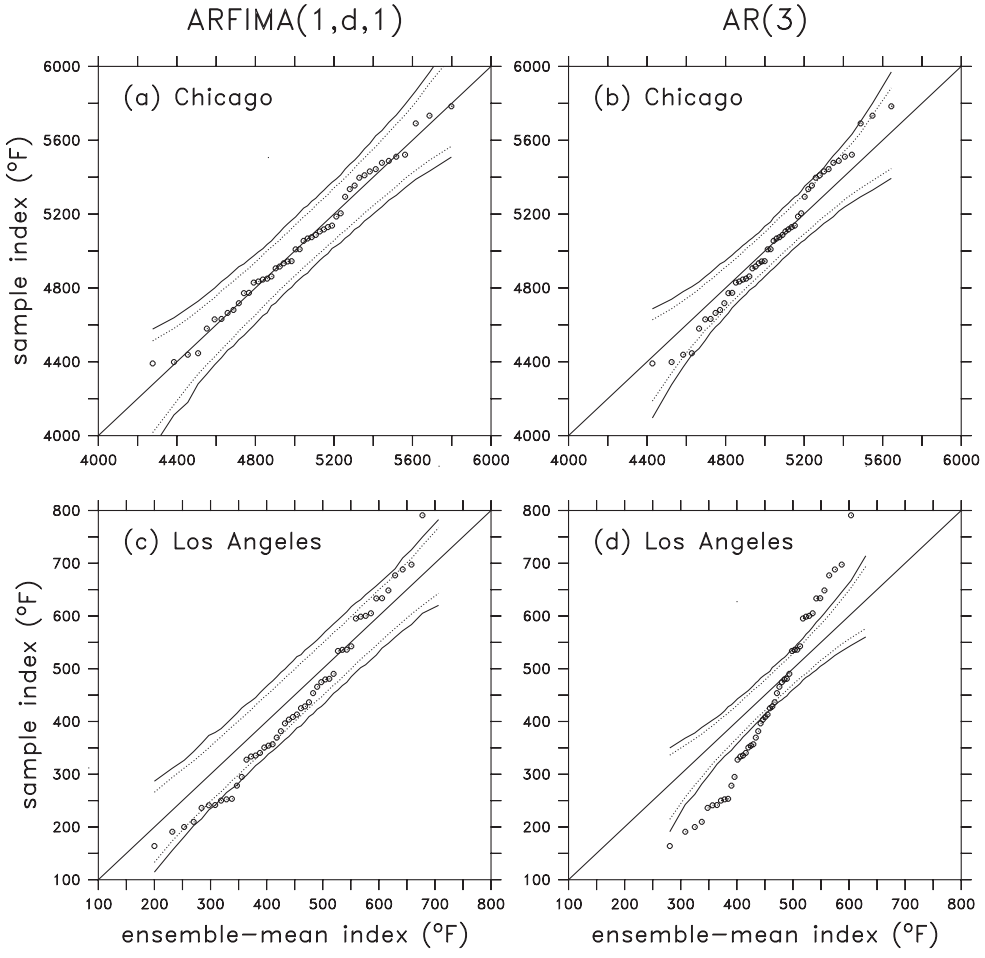

Fig. 5. *q*-*q* plots of the observed against the modeled index quantiles (dots) for (a,b) Chicago and (c,d) Los Angeles using the ARFIMA(1,*d*,1) model (a,c) and the AR(3) model (b,d). Solid (dotted) curves show 99% (95%) confidence envelopes. The central solid line is a diagonal for comparison. Units are degree days in Fahrenheit

The Monte Carlo test can also be applied non-parametrically by computing a sample cumulative distribution function (CDF) for each of the 50-member subsamples and estimating confidence intervals for each of the CDF quantiles. The results of this procedure are displayed in *q*-*q* form in Fig. 5. In this figure, the dots indicate (*x*,*y*) pairs, where *x* is the model population-mean quantile and *y* the historical data sample quantile. If the 2 CDFs were identical, the dots would lie along the diagonal. The solid (dotted) lines above and below the diagonal indicate 99% (95%) confidence intervals for each quantile. The figure confirms the results obtained above. The ARFIMA model generally does better than the AR at both stations, the dots being more closely aligned with the diagonal. However, both models fail in Los Angeles, with several dots exiting the 99% confidence envelope and a pronounced systematic difference between the modeled and observed CDFs.

5.5. Seasonality in the autocorrelation structure and its effects

We saw in the previous section that while the ARFIMA(1,*d*,1) model generally performs better than the AR(3), it still fails in Los Angeles, despite the close fit to the observed autocorrelation function displayed in Fig. 3b. It turns out that one major reason for this is that the model was fitted to the entire data series. This is only correct if the autocorrelation of the data is stationary. As Fig. 6 clearly shows, this is not the case. The curves in the figure were obtained by computing the autocorrelation function separately for each summer (winter) and then averaging over all summers (winters). There is clearly much greater persistence during summer than winter. The physical reasons for this are not clear; we may speculate that during summer the high-frequency 'weather' activity is much lower, so that the low-frequency variability (plausibly of oceanic origin,

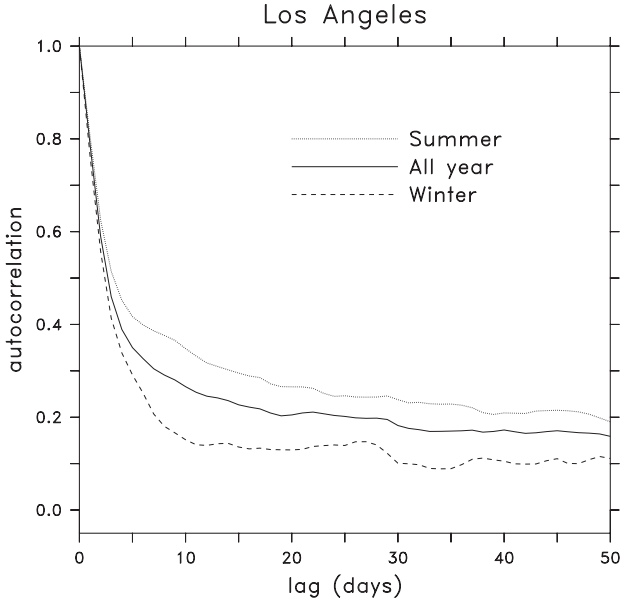

Fig. 6. Seasonal variation of the autocorrelation structure of observed SAT in Los Angeles: summer and winter average autocorrelations. The all-year autocorrelation function is replotted from Fig. 3b

given the coastal location of this station) accounts for a greater fraction of the total variance.

From our applied point of view, the consequence is that if we fit the model to the 'all year' data set, we will considerably underestimate the autocorrelation during summer and hence, by Eq. (11), the variance of a summer-based index. This is exactly what is observed in the results of the previous section. We note that in Chicago the autocorrelation structure varies much less from season to season (not shown), so the issue is not so critical there.

A simple 'fudge' to correct matters is to fit the model separately to summer data. This was done by extracting summer (May–September) data from each year, and then fitting the model to the single time series obtained by juxtaposing all the summers. The results of this operation are displayed in Fig. 7. We see that the ARFIMA model now performs reasonably well, with only 1 point in the tail of the distribution falling outside the 99% confidence envelope. The AR(3) model, on the other hand, continues to perform poorly, giving a significant underestimate of the index variance. In any event, we stress that this is only a temporary solution; a more satisfactory one would be to fit the entire data series using seasonally varying parameters. We are currently investigating techniques for doing this.

6. SOME REMARKS ON THE ORIGIN OF LONG MEMORY IN SAT

In Section 1, we saw that simple physical considerations based on the behaviour of mid-latitude synoptic-scale eddies led to the Ornstein-Uhlenbeck process, which successfully accounts for the high-frequency behaviour of SAT but underestimates the low-frequency variance. We then showed (Section 4) that ARFIMA models can successfully capture both the high- and low-frequency behaviour. However, ARFIMA models are mathematical tools which, though useful for applications, have no immediate physical interpretation. That is, they do not provide a direct

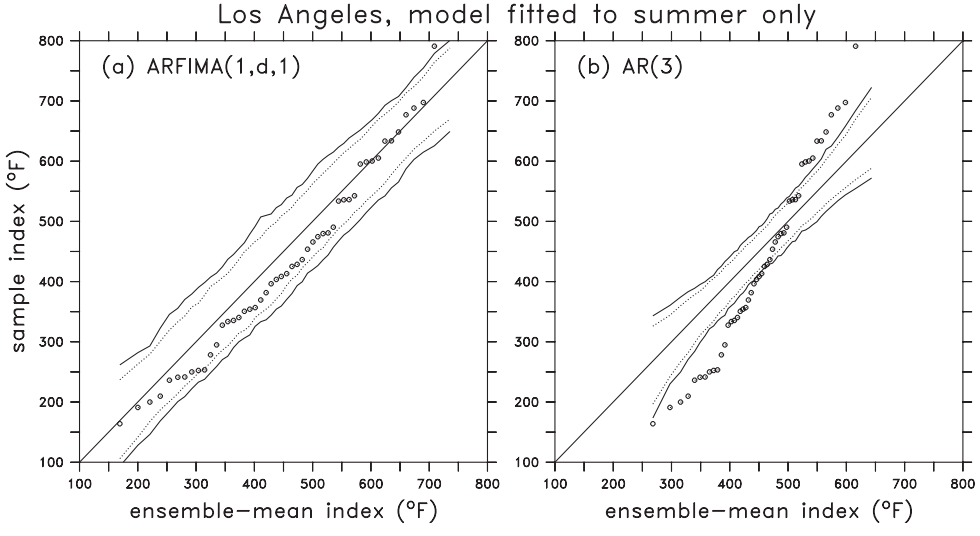

Fig. 7. *q-q* plots of the observed against the modeled index quantiles (dots) for Los Angeles using the (a) ARFIMA(1,*d*,1) and (b) AR(3) models fitted to summer data only

answer to the basic physical question of why there is long memory in SAT time series.

The adequacy of the Ornstein-Uhlenbeck process at high frequency suggests we are on the right track in modeling the effect of high-frequency transients. It is then plausible to attribute the 'extra' variance at low frequency to other, slower processes which also affect SAT. The identification and analysis of mechanisms for the low-frequency atmospheric variability has been a central theme of climate research over the past 2 decades, and is still ongoing. A few of the candidates are regime-like behaviour in the atmosphere (Hansen & Sutera 1995), local or remote oceanic forcing (Wallace & Gutzler 1981), and atmosphere-land surface interaction (Manabe & Stouffer 1996).

How can we incorporate the slow mechanisms into the model? The simplest option (Granger 1980) is to assume that each mechanism affecting SAT can be modeled as an AR(1) process with a certain value of á and that all mechanisms act independently of each other. The final SAT process is then just the sum of a certain number N of AR(1) processes:

$$x_i = \sum_{n=1}^{N} y_i^{(n)} \qquad (12)$$

where $y_i^{(n)}$ are the individual AR(1) processes. It can be shown that, if the coefficients of the AR(1) processes are appropriately chosen, then as $N \to \infty$ the aggregate process converges to one having the long-memory property (Eq. 4).

In the real atmosphere there is presumably only a finite number of relevant processes affecting SAT. Thus, $N < 8$, and the aggregate time series will not strictly speaking have the long-memory property: the spectrum will not behave as a power law all the way to the origin but will eventually flatten out to a constant. However, the cross-over may occur at very low frequency and may not be detectable given the finite-length data series available. We illustrate this point with a specific example. We consider 3 independent AR(1) processes with parameter values 0.82, 0.95 and 0.999 (corresponding to time scales of 6, 20 and 1000 d) and noise variances 1, 0.3 and 0.01 respectively. We generate 105 d long series with each model, compute the aggregate series (12) and take its power spectrum. The result, shown in Fig. 8a, has the same qualitative features as the observed spectra reported in Fig. 1: a high-frequency part with slope close to −2 crossing over to a shallower slope at low frequencies. The periodogram test (Section 3.1) applied to this series indicates that long memory is present with intensity d = 0.12 ± 0.07. Note that the process parameters employed here have been selected arbitrarily for illustrative purposes only. It may be possible to devise an appropriate parameter-selection algorithm giving an optimal fit to any given time series, but we do not pursue this issue here.

One might argue that it is unrealistic to assume that the various processes affecting SAT occur independently of each other. In fact, a large part of the variability is actually *generated* by the coupling of the various parts of the climate system. For instance, much of the variability in mid-latitude oceanic temperatures can be attributed to stochastic forcing by atmospheric transients (Frankignoul 1995). The simplest way to model such interacting processes is with a multivariate AR(1) model:

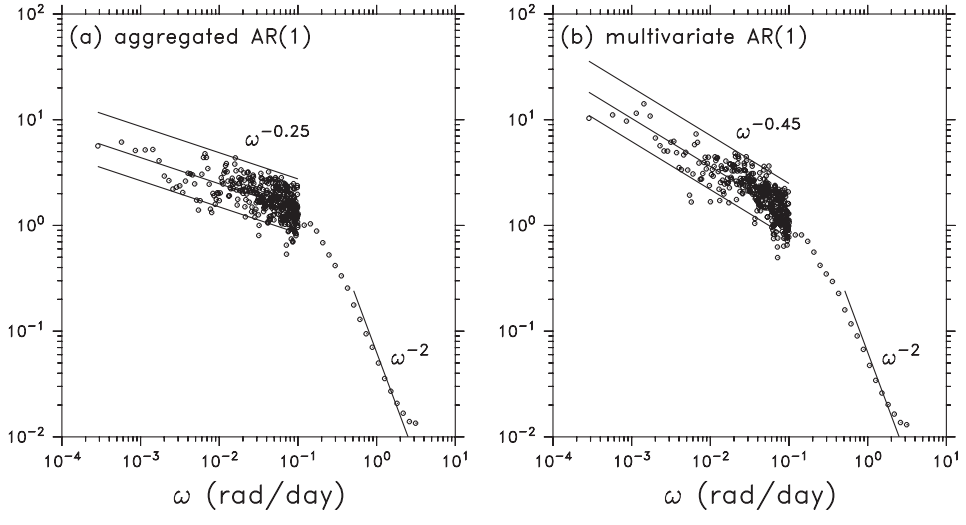

Fig. 8. Power spectral density estimates for synthetic data produced by (a) the aggregation of 3 independent AR(1) processes and (b) a trivariate AR(1) process. See Fig. 1 and Section 6 for details

$$\boldsymbol{x}_i = \mathbf{A}\boldsymbol{x}_{i-1} + \mathbf{B}\boldsymbol{e}_i \quad (13)$$

where $\boldsymbol{x}_i = (x^1_i, x^2_i, \ldots, x^n_i)$, $\boldsymbol{e}_i = (\varepsilon^1_i, \varepsilon^2_i, \ldots, \varepsilon^n_i)$, is a vector of n independent unit-variance white noise processes and **A** and **B** are $n \times n$ matrices. Let us again consider a specific example. We take:

$$\mathbf{A} = \begin{pmatrix} 0.82 & 0.25 & 0.085 \\ 0.03 & 0.9 & 0.0 \\ 0.001 & 0.0 & 0.998 \end{pmatrix} \quad \mathbf{B} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad (14)$$

Only x^1, which we may think of as representing high-frequency atmospheric transients, is directly forced by external white noise. Components x^2 and x^3 are essentially AR(1) processes with long decay time scales, stochastically forced by the 'atmosphere' x^1; we may think of them as representing slow components of the climate system such as the land surface and the ocean. These slow components in turn feed back onto the atmosphere through the top-row elements in **A**. The particular values of the entries in **A** are entirely arbitrary and serve only for illustration. We generate a 105 d long run of the process and compute the power spectrum of the atmospheric component only. The result (Fig. 8b) again looks quite similar to observations. The periodogram test gives $d = 0.23 \pm 0.07$. Again, any value of d is obtainable by manipulating **A**.

7. SUMMARY AND DISCUSSION

We have analysed 3 multidecadal daily SAT time series representative of conditions in the northern mid-latitudes. We have applied a number of tests to detect the presence of long memory; these indicate that long memory is indeed present in all 3 time series. We have shown that an ARFIMA(1,d,1) model gives a qualitatively unbiased fit to the autocorrelation structure of the data out to lags of a season or more, while AR models even of high order give a considerable underestimate of the high-lag autocovariance. For the reasons outlined in Section 5.3, ARFIMA models are therefore much better suited to weather-derivative pricing than AR models employing a similar number of parameters.

Finally, we have shown (Section 6) how spectra with the same qualitative features as those of observed SAT (Fig. 1) can be generated by simple aggregation of several short-memory processes which may be independent or coupled. We note that, in the coupled case, our argument is very close to the stochastic climate model of Hasselmann (1976). The reasoning behind the stochastic climate model is as follows: The atmosphere, if left to its own devices, will produce a spectrum which is white for time scales longer than about 10 d. Atmospheric variability will act as a source of stochastic forcing on the ocean, due to fluctuations in wind stress and heat flux. Sea-surface-temperature (SST) perturbations have a typical decay rate on the order of several months. This scale separation allows the atmospheric forcing of the ocean to be modeled as a white noise. Thus the SST can be modeled as an Ornstein-Uhlenbeck process with intrinsic time scale of several months. In Section 6, we suggested that the variability induced in the ocean may in turn feed back onto the atmosphere and contribute to its low-frequency variability, thus reddening the low-frequency tail of the atmospheric spectra. While this is quite plausible in the Tropics, where atmosphere-ocean coupling is strong, it is less obvious in the mid-latitudes, where the coupling is much weaker. Feedback of mid-latitude SST variability onto the atmosphere has, however, been documented in a GCM (general circulation model) by Rodwell et al. (1999). This explanation for the appearance of long memory in atmospheric time series is similar to that suggested in Tsonis et al. (1999). Other explanations, relying on internal dynamics of the atmospheric boundary layer, have been suggested in the literature (Jánosi & Vattay 1992, Pelletier 1997). Further work is needed to decide among these alternatives.

We summarise our main conclusions as follows:

- Long memory of moderate intensity ($d \sim 0.1$ to 0.25) can be detected in the 3 mid-latitude SAT time series studied here with 95% statistical confidence;
- A simple explanation for the apparent presence of long memory in these time series is that SAT is simultaneously affected by a number of physical processes, each of short memory but with widely disparate intrinsic time scales;
- ARFIMA models with only 3 parameters give an excellent fit to SAT time series;
- ARFIMA models are suitable for pricing weather derivatives, provided care is taken to account for seasonality in the autocovariance structure.

Acknowledgements. R.C. was supported by Dansk Grundforskningsfond. The Central England temperature data were obtained through the Data Support Section of the National Center for Atmospheric Research (http://dss.ucar.edu/datasets/ds825.0/data). The 'R' environment is freely available at http://cran.stat.wisc.edu. All figures were prepared using Ferret software, freely available at: http://ferret.wrc.noaa.gov.

LITERATURE CITED

Beran J (1989) A test of location for data with slowly decaying serial correlations. Biometrika 76:261–269

Beran J (1994) Statistics for long-memory processes. No. 61, Monographs on Statistics and Applied Probability. Chapman & Hall, New York, CRC, Boca Raton, FL

Bloomfield P (1992) Trends in global temperature. Clim Change 21:1–16

Box GEP, Jenkins GM (1970) Time series analysis, forecasting and control. Holden-Day, San Francisco
Cotton WR, Pielke RA (1995) Human impacts on weather and climate. Cambridge University Press, Cambridge
Frankignoul C (1995) Climate spectra and stochastic climate models. In: von Storch H, Navarra A (eds) Analysis of climate variability: applications of statistical techniques. Springer Verlag, Berlin, p 139–157
Granger CWJ (1980) Long memory relationships and the aggregation of dynamical models. J Econometr 14: 227–238
Granger CWJ, Joyeux R (1980) An introduction to long-range time series models and fractional differencing. J Time Ser Anal 1:15–30
Hansen AR, Sutera A (1995) The probability density distribution of planetary-scale atmospheric wave amplitude revisited. J Atmos Sci 52:2463–2472
Haslett J, Raftery AE (1989) Space-time modelling with long memory dependence: assessing Ireland's wind power resource. Appl Stat 38:1–50
Hasselmann KF (1976) Stochastic climate models. Part I: Theory. Tellus 28:473–484
Hosking JRM (1981) Fractional differencing. Biometrika 68: 165–176
Hull JC (1998) Introduction to futures and options markets, 3rd edn. Prentice-Hall International, New York
Hurst HE (1951) Long-term storage capacity of reservoirs. Trans Am Soc Civil Eng 116:770–799
Jánosi IM, Vattay G (1992) Soft turbulent state of the atmospheric boundary layer. Phys Rev A 46:6386–6389
Jewson S (2000) Use of GCM forecasts in financial-meteorological models. In: Proceedings of the 25th Annual Climate Diagnostics and Prediction Workshop. US Department of Commerce, Washington, DC
Katz RW, Parlange MB (1998) Overdispersion phenomenon in stochastic modeling of precipitation. J Clim 11:591–601
Koscielny-Bunde E, Bunde A, Havlin S, Goldreich Y (1996) Analysis of daily temperature fluctuations. Physica A 231: 393–396
Manabe S, Stouffer RJ (1996) Low-frequency variability of surface air temperature in a 1000-year integration of a coupled atmosphere-ocean-land surface model. J Clim 9: 376–393
Mandelbrot BB, Wallis JR (1969) Some long-run properties of geophysical records. Water Resour Res 5:321–340
Montanari A, Rosso R, Taqqu MS (1996) Some long-run properties of rainfall records in Italy. J Geophys Res 101: 29431–29438
Parker DE, Legg TP, Folland CK (1992) A new daily Central England temperature time series, 1772–1991. Int J Climatol 12:585–596
Peixoto JP, Ort AH (1992) Physics of climate. American Institute of Physics, New York
Pelletier JD (1997) Analysis and modeling of the natural variability of climate. J Clim 10:1331–1342
Peng CK, Buldyrev SV, Goldberger AL, Havlin S, Sciortino S, Simons M, Stanley HE (1992) Long-range correlations in nucleotide sequences. Nature 356:168–170
Peng CK, Buldyrev SV, Havlin S, Simons M, Stanley HE, Goldberger AL (1994) Mosaic organization of DNA nucleotides. Phys Rev E 49:1685–1689
Rodwell MJ, Rowell DP, Folland CK (1999) Oceanic forcing of the wintertime North Atlantic oscillation and European climate. Nature 398:320–323
Shea DJ, Madden RA (1990) Potential for long-range predictability of monthly-mean surface temperatures over North America. J Clim 3:1444–1451
Simmons AJ, Hoskins BJ (1978) The life cycles of some nonlinear baroclinic waves. J Atmos Sci 35:414–432
Stephenson DB, Pavan V, Bojariu R (2000) Is the North Atlantic Oscillation a random walk? Int J Climatol 20:1–18
Syroka J, Toumi R (2001) Scaling and persistence in observed and modeled surface temperature. Geophys Res Lett 28: 3255–3258
Taqqu MS, Teverovsky V, Willinger W (1995) Estimators for long-range dependence: an empirical study. Fractals 3: 785–798
Teverovsky V, Taqqu MS (1997) Testing for long-range dependence in the presence of shifting means or a slowly declining trend using a variance type estimator. J Time Ser Anal 18:279–304
Tsonis AA, Roebber PJ, Elsner JB (1999) Long-range correlations in the extratropical atmospheric circulation: origins and implications. J Clim 12:1534–1541
von Storch H, Zwiers FW (1999) Statistical analysis in climate research. Cambridge University Press, Cambridge
Wallace JM, Gutzler DS (1981) Teleconnections in the geopotential height field during the northern hemisphere winter. Mon Weather Rev 109:784–812
Zeng L (2000) Weather derivatives and weather insurance: concept, application and analysis. Bull Am Meteorol Soc 81:2075–2082

Editorial responsibility: Hans von Storch, Geesthacht, Germany

Submitted: May 24, 2001; Accepted: October 21, 2001
Proofs received from author(s): May 15, 2002