Australian Meteorological and Oceanographic Journal 63 (2013) 473–486

A stochastic model for runs of extreme days for a daily meteorological variable

Warwick Grace
Grace Research Network

(Manuscript received February 2013; revised September 2013)

The expected frequency of runs of extreme days is modelled with a seasonal autoregressive representation using two site-specific scalar parameters—autocorrelation and an index of seasonality. Extreme days for a meteorological variable such as temperature, evaporation, sunshine and wind run are defined as days when the variable in question exceeds an arbitrary upper threshold or fails to exceed a lower threshold. The model estimates the frequency of runs as a function of duration (in days) and percentile or absolute threshold.

The model is applied to several types of daily variable such as wind run, evaporation, sunshine and pressure, and tested extensively on datasets of daily maximum and minimum temperature for 60 to 70 sites in each of Australia and Europe. Using a priori parameter values of autocorrelation and seasonality, the model often provides fair agreement with the observed frequency of runs of extreme days. With the parameter values tuned to fit the observed frequency of runs, the agreement is typically fair to good. It is concluded that the model has the potential to estimate frequency (or return period) of unusually long runs with very low or high percentile thresholds.

Introduction

Extremes of many meteorological variables are of interest because of their impact on society and the environment. An extreme day is here defined as one where the variable concerned exceeds some chosen upper threshold or fails to exceed a chosen lower threshold; and a run of extreme days is a sequence of consecutive extreme days. To avoid ambiguity regarding the duration, the duration of a run is equal to the maximum number of extreme days in the run. Most studies of climate extremes have focussed on temperatures or rainfall: for instance, the World Meteorological Organization Commission for Climatology/Climate Variability and Predictability/Joint Technical Commission for Oceanography and Marine Meteorology Expert Team on Climate Change Detection and Indices (ETCCDI) have 27 indices of extremes most of which are related to temperature or rainfall, either as single events or as a run of such events (see http://etccdi.pacificclimate.org/list_27_indices.shtml).

However, the stochastic model to be presented here is intended to be general and applicable to any meteorological variable which is measured daily and which exhibits an annual cycle. The model's output—the frequency of runs of extreme days—is not necessarily the best metric to typify spells of extremes: for example, although the model may be applied to the temperature record to model the frequency of runs of extremely hot days, there are other more suitable metrics for heatwaves (Perkins and Alexander 2013).

Stochastic modelling of runs of extreme days has been mostly related to heatwaves and wet or dry spells. Stochastic time series modelling using first-order autoregressive (AR) models with a knowledge of the monthly mean, standard deviation and autocorrelation provide characteristics of heatwaves that are in good agreement with observations in mid-latitude areas (Mearns et al. 1984, Kysely 2010, Grace 2011). Grace (2011) found that an analytical Markov model with one empirical coefficient was comparable to the AR model. The theory of AR models is described and illustrated with several climate examples in the texts of von Storch and Zwiers (1999) and Wilks (2011). Stochastic modelling of runs of wet and dry days (not specifically for extremes) and synthetic weather generation has been performed using Markov models (Wilks 2008, Wilks 2011) and these too usually require separate parameters for each month.

There appears to be no simple analytical or parametric model of frequency of runs of extremes of a daily variable in the literature other than the analytically expressed Markov model of Grace et al. (2009) and Grace (2011). The purpose of this paper is to present an autoregressive model

Corresponding author address: Warwick Grace, Grace Research Network, 29 Yurilla Drive, Bellevue Heights, 5050, Australia. Email: wg@graceresearch.com, phone: +61 8 8277 6847

Table 1. Mean performance measures of correlation coefficient r, the χ^2 test, and $RMSEA$ (see p478) for each dataset for runs of days with maximum temperatures above the 90th, 95th and 98th percentiles and minimum temperatures below the 10th, 5th and 2nd percentiles for Australian and European datasets for years up to 1970. Australian (European) dataset was restricted to sites with at least 25 (70) years of continuous record. F is percentage of sites for which null hypothesis is accepted, under the χ^2 test at 0.05 significance level. $RMSEA$ values up to 0.05 indicate 'close approximate fit' and up to 0.08 'reasonable approximate fit'. Measures for the untuned model are shown in normal type, and those for the tuned model in bold.

Dataset	*Sites*							
For runs of maximum temperatures above the 90th, 95th and 98th percentiles								
		r		*F %*		*RMSEA*		
Australian up to 1970	67	0.95	**0.95**	29	**68**	0.07	**0.04**	
European up to 1970 with ≥ 70 years	65	0.93	**0.94**	22	**70**	0.07	**0.03**	
Overall 'heatwaves'	132	0.94	**0.94**	25	**69**	0.07	**0.04**	
For runs of minimum temperatures below the 10th , 5th and 2nd percentiles								
		r		*F %*		*RMSEA*		
Australian up to 1970	70	0.96	**0.96**	21	**70**	0.11	**0.05**	
European up to 1970 with ≥ 70 years	62	0.93	**0.94**	11	**37**	0.12	**0.07**	
Overall 'cold spells'	132	0.95	**0.95**	16	**54**	0.11	**0.06**	
Overall mean for 'heat waves' and 'cold spells'	264	0.94	**0.95**	22	**61**	0.08	**0.05**	

of runs frequency which is effectively parametric in that it is fully described by a few parameters (two in the case of this model) and is general enough to apply to many typical weather variables over a wide range of designated threshold percentiles and run durations. The testing datasets are described then the model theory developed. The model is then applied and model performance assessed qualitatively and quantitatively. Discussion and conclusions follow.

Data considerations

For wind run, evaporation, pressure and sunshine hours, all available daily data as at July 2011 (Bureau of Meteorology 2011) for Adelaide Airport were used to investigate the model's performance. Adelaide Airport was used in preference to the city's near-CBD site to avoid possible trends from urbanisation effects.

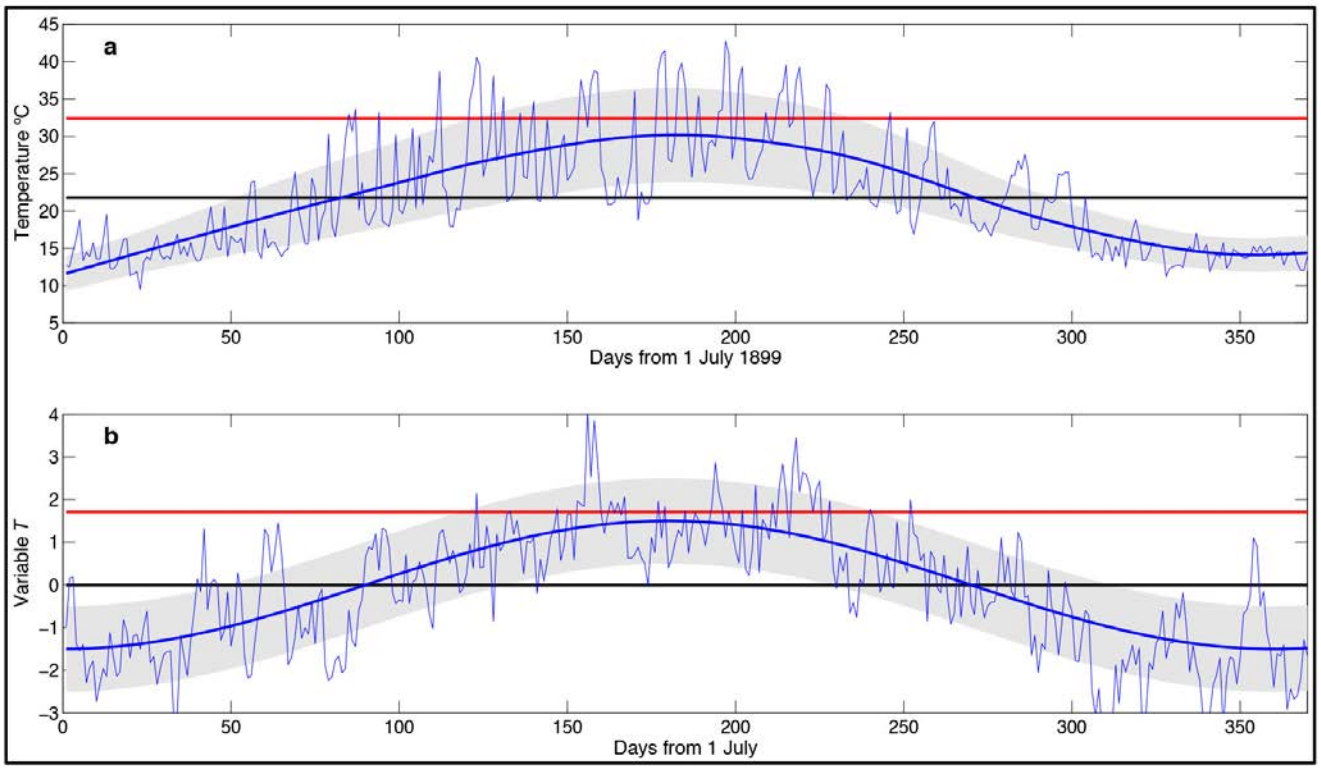
Other data used are the daily maximum and minimum temperature record from the Australian Bureau of Meteorology's ACORN-SAT High Quality Temperature dataset (Bureau of Meteorology 2012, Trewin 2013) and the quality-controlled daily maximum and minimum temperature dataset for sites in Europe, including Russia, and the Mediterranean available from the European Climate Assessment Dataset (Klein Tank et al. 2002 and Klok and Klein Tank 2009). The data were allocated to either calendar or 'Austral summer' years (from July to June) depending as to what time of year the extremes in question tend to occur. A complete, or near-complete, year of record at a station is regarded as one with no more than two missing observations. For the years with one or two missing observations, the missing daily observations were substituted with linearly interpolated values. Only complete or near-complete years were used.

Fig. 1. Schematic of a sequence of days with three runs of extreme days. Extreme days are represented as grey. From the left, the three runs are of length one, three and two days.

For simplicity the confounding effect of global warming was avoided by using only years up to 1970 in the temperature datasets. The period of up to 1970 was chosen as being approximately stationary since it is known that most of the warming trend in maximum temperatures in Australia in the recent past has occurred since 1970 (Commonwealth of Australia 2012). For the Australian dataset, with two exceptions noted later, only sites with at least 25 years of complete or near-complete data up to 1970 were used. For the European dataset, only sites with at least 70 years of complete or near-complete data up to 1970 were used. The numbers of sites are shown at Table 1.

Temperature-percentile relationships were constructed for each site for the whole period of record (up to 1970) and calculated over the entire record regardless of the annual cycle. Thus it is always straightforward to convert runs in terms of percentile thresholds to runs in terms of absolute temperatures. For example, the 90th percentile maximum temperature for Melbourne for the period up to 1970 was 29.4 °C while 30 °C corresponds to 91.2th percentile.

Fig. 2. (a) A typical annual cycle of daily maximum temperature (thin black line) for Adelaide (West Terrace). The period shown is 1 July 1899 to 30 June 1900. Day 183 is 1 January 1900. The mean (thick black line) and 90th percentile (thick red line) are computed from the 30-year period 1887–1916. The grey shading presents ±1 standard deviation about the seasonal cycle (thick blue line), computed over the period 1887–1916. (b) A simulation using Equation 2 with an overall mean of 0, a standard deviation of 1.0, an autocorrelation of 0.6 and a sinusoidal seasonal variation of the mean with an amplitude equal to 1.5. For both observations and the simulation, occurrences above the 90th percentile tend to occur in and around summer or its equivalent.



Model description and theory

Assumptions and definitions

For a variable T of a meteorological nature, a stationary climate is assumed, and taken to mean cyclostationary, that is, stationary after seasonal adjustment (Wilks 2011). An extreme day is one where T exceeds (or remains below) a specified upper (or lower) threshold, T_p. The threshold T_p is the p^{th} percentile value calculated over the entire record regardless of the annual cycle. Extreme days and non-extreme days are mutually exclusive, and a run is a sequence of extreme days bounded by non-extreme days (see Fig. 1).

Stochastic simulation

Autoregressive models of order one have been used to model heatwaves in mid-latitudes (Mearns et al. 1984, Kysely 2010, Grace 2011). Von Storch and Zwiers (1999) and Wilks (2011) provide mathematical theory and examples of climate applications of stochastic AR models. In this paper the focus is on an idealised stationary time series with constant standard deviation and autocorrelation and with a seasonal mean of a sinusoidal nature.

The underpinning idea of the idealisation is shown schematically at Fig 2. The assumption is that only two parameters are sufficient to capture the essential features of the real observations. The two parameters are autocorrelation and an index of seasonality which is the amplitude of the seasonal cycle scaled by the standard deviation. To illustrate this, Fig. 2 shows the time series of daily maximum temperature for the year 1899–1900 at Adelaide (West Terrace city site) compared to an idealised simulation with standard deviation set to unity and with two constant parameters—daily autocorrelation of 0.6 and a seasonal sinusoidal variation of amplitude 1.5 standard deviation units with these parameters being close to values calculated in the manner detailed later. Firstly, the observed seasonal cycle is approximately sinusoidal. Secondly, extremes (those days exceeding 90th percentile maximum temperature in this illustration) tend to occur in and around a peak season. Although both the observed standard deviation and the autocorrelation vary throughout the year (not shown), the seasonal shift in the mean reduces or eliminates their importance in producing extremes outside of the summer period. Nevertheless extremes can still occur outside of peak season both in the observations and in the simulation. Overall, the three most important parameters are assumed to be the autocorrelation in peak season, the standard deviation in peak season and the seasonal amplitude of the mean. It will be shown later that the last two parameters can be combined to form a dimensionless

seasonality index so that only two parameters are required for the simulation.

The assumption that a time series conforms to an AR(1) model is articulated as (Wilks 2011):

$$T_{m+1} = (1-a)\mu + aT_m + \sqrt{(1-a^2)}\sigma\phi(0,1), \quad ...(1)$$

where T_m represents the daily variable at day m = 0, 1, 2, 3,…, M, and a is the autocorrelation at lag 1, μ is the mean of the variable T, ϕ is a standard normally distributed random (noise) variable and σ is the standard deviation of T about μ. We begin with the simpler case where μ is held constant. Without loss of generality we can set $\mu = 0$ and $\sigma = 1$. A set of simulations each covering 100 000 years (M = 36 500 000 days) is undertaken. To initialise the simulations, T_0 is set equal to the mean, which here is zero.

From these simulations R is estimated as the number of occurrences per century for runs of length x = 1, 2, 3,… days when the variable exceeds thresholds of p = 75, 76, 77,…98, 99th percentiles. In other words, R is a function of run length (x days) and the percentile threshold, p, for a given autocorrelation, a. Von Storch and Zwiers (1999, their Fig. 10.8) present a similar example of runs frequency against run length for a = 0, 0.3, and 0.9 but using a threshold at the 50th percentile.

Idealised seasonality

A more realistic regime is provided by the inclusion of a seasonal variation of the mean. Here the model continues to have constant autocorrelation a, constant standard deviation ($\sigma = 1$) but the mean μ varies sinusoidally over the annual cycle with amplitude of s about a long term mean of zero. Thus the mean μ oscillates about zero from $-s$ in off-peak to $+s$ in mid-peak and is set equal to $-s\cos(2\pi m/365)$ where s is essentially a multiple of the standard deviation which remains set at 1.

So the daily variable is simulated stochastically with two inputs, a and s, by Equation 2:

$$T_{m+1} = -(1-a)s\cos(2\pi m/365) + aT_m + \sqrt{(1-a^2)}\phi(0,1). \quad ...(2)$$

The start and end of the series will be during the off-peak period (when m = 0 or a multiple of 365, the first right-hand term will be at its lowest) so edge effects from incomplete runs at the start or end are avoided. As before, from these simulations R is tabulated in relation to x and p for any combination of a and s which allows the model to be expressed as

$$R = R(x, p; a, s) \quad ...(3)$$

where the semicolon indicates that x and p are regarded as variables in the usual mathematical sense and the a and s terms are parameters. R is a family of probability functions scaled by the number of days in a century. $100/R$ may be interpreted as the return period in years. Equation 3 is a parametric function in the sense that a family of curves is described by some mathematical procedure involving (in this model, two) parameters and the family of curves either 'predicts' observed outcomes or the parameters can be set so that the family of curves approximates empirically observed outcomes.

In estimating the runs frequency an alternative algorithm is to simulate a daily sequence of 100 years and then to average the results of 1000 such sequences. The potential advantage is that a secular trend over the 100 year period could be incorporated. This would then enable a comparison of runs frequency between the early decades and the later decades of a century subject to a warming trend.

Although Equation 2 is stochastic, if M is sufficiently large then in practice R is a deterministic function. Obtaining R from the stochastic algorithm is computationally intensive but the disadvantage may be offset by the one-time creation of a look-up table. This was done by nesting incrementing values of a and s and conducting separate simulations to determine R for suitable ranges of x and p. The resulting look-up table is four-dimensional in x, p, a and s and capable of interpolation. A diagrammatic representation in Fig. 3 of R, as a family of curves in x and p for nine selected value-pairs of a and s, shows the response as autocorrelation and/or seasonality vary. On these diagrams, the vertical axes are logarithmic. For small a and/or s the model lines are steeper and straighter. The simplest case where both a and s are zero reduces to a series of independent Bernoulli events and the model lines are expected to be exponential, that is, to plot as straight lines on a log-linear graph (see Appendix A). As a and/or s increase for a given p, the lines become less steep implying that longer runs of extremes occur more frequently and shorter runs occur relatively less frequently. Also the lines develop a concave upward curvature with increasing a and/or s implying that an additional feature to the exponentiality is involved.

Estimating the model parameters (a priori)

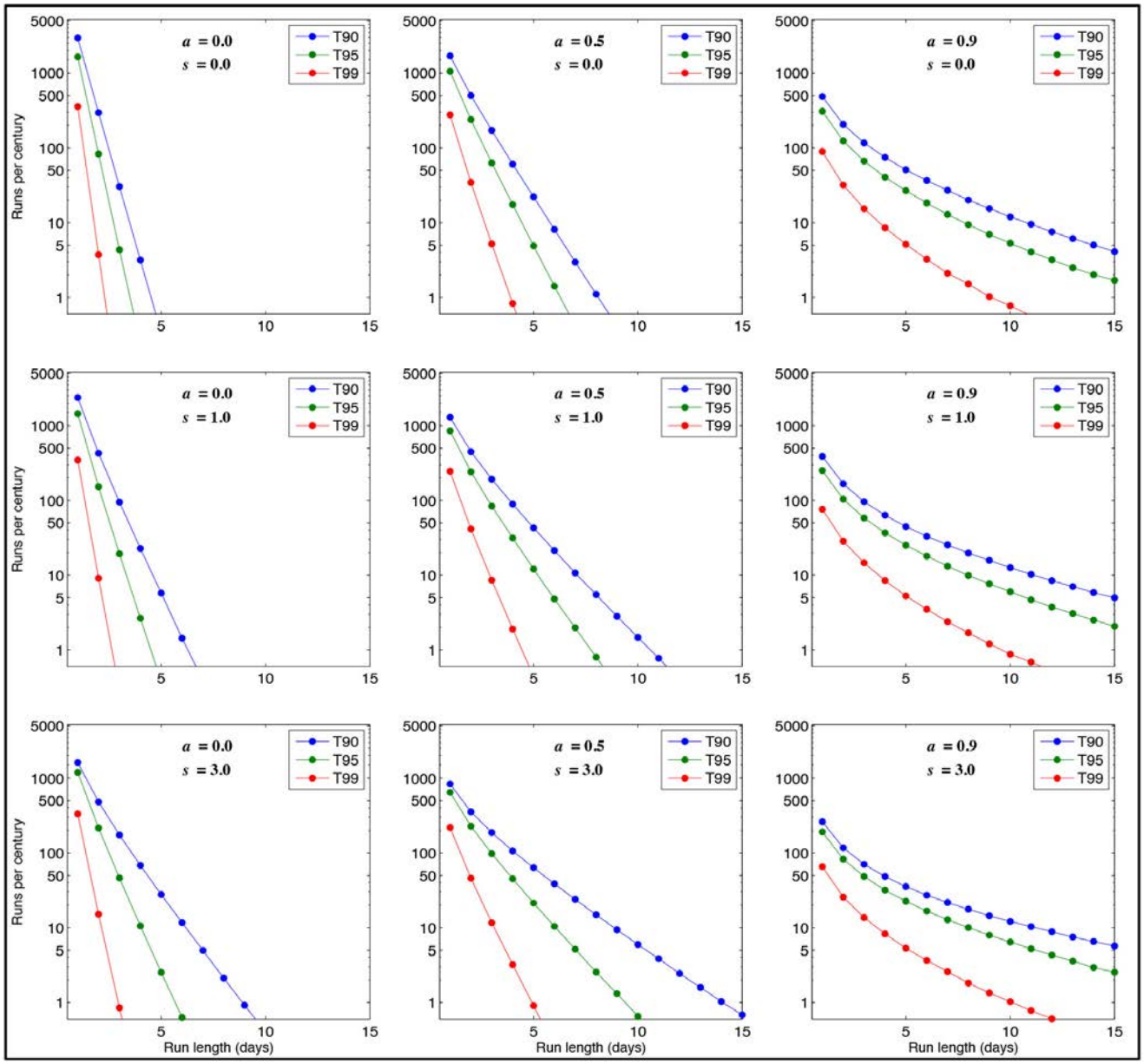
Although the autocorrelation varies on a seasonal basis, it is intuitively apparent that the autocorrelation in the peak season is most relevant since outside of the peak season the extremes are much less likely to occur. To simplify the discussion, the notional peak period is assumed to be the month of January. Sample autocorrelations (at lag of one day) are calculated for each January; the mean of these is taken as the estimate of autocorrelation for the peak period.

For a measure of seasonality we use the following non-dimensional index

$$s = \frac{T_{Jmean} - T_{Amean}}{\sigma_J} \quad ...(4)$$

where T_{Jmean} is the mean during the peak month of January, T_{Amean} is the mean over the whole year and σ_J is the standard deviation of the daily values during January. As with the estimation of autocorrelation, s is estimated as the mean of the individual January values of s. A value of s = 0 corresponds to a seasonless regime while large s implies a short peak period. If the standard deviation were

Fig. 3. Model curves for selected pairings of autocorrelation a and seasonality s shown by plots of R against runs duration in days x with selected thresholds at percentiles p (top to bottom: 90, dark blue; 95, green; 99 per cent, red).



constant over the seasonal cycle and the monthly means followed a sine curve then the expression for s is equivalent to the seasonality measure used in the idealised regime above. The calculation of the parameters is still practical for incomplete datasets. It is also noted that a secular trend in T would not affect the calculation of a or s.

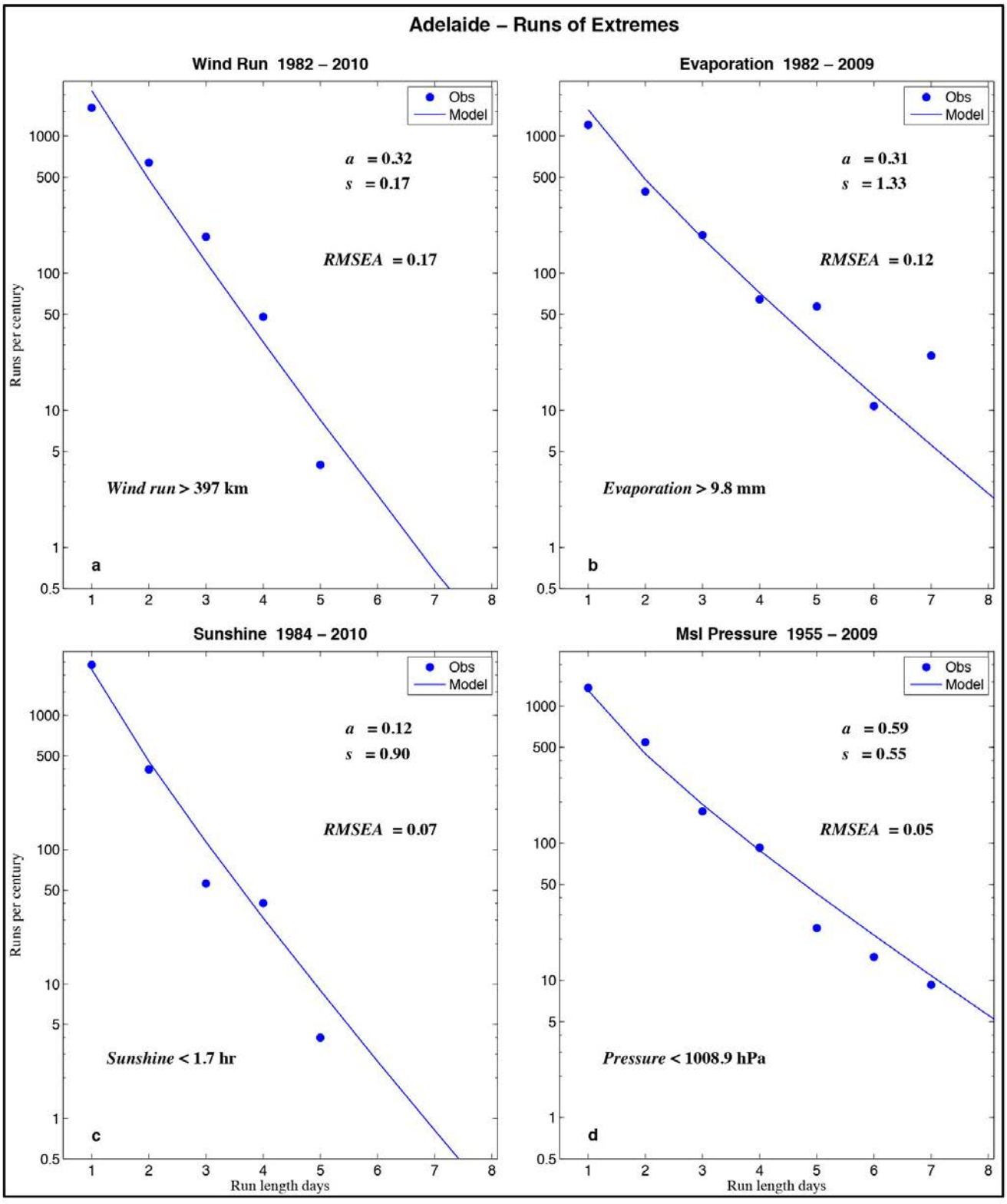
Bias adjustment for autocorrelation and standard deviation

The estimation of autocorrelation and standard deviation is as the mean of the sample estimates as described above. However, autocorrelation in a series has the effect of biasing the sample estimates of moments (other than the mean) and of the autocorrelation itself (see Appendix B). To de-bias the sample estimates of standard deviation and autocorrelation, the expected bias needs to be taken into account. The expected bias can be determined from the sample autocorrelation and the sampling size (in this case 30) and the procedure to do this is detailed in Appendix B. In the model assessment process it was found that neglecting the bias usually caused apparent deterioration in model performance.

Estimating the model parameters (a posteriori)

A second approach to estimate the parameters a and s is to compare the model output of runs frequencies to observations and find the best fit of model to observations. These tuned, or effective, values are denoted as a^* and s^* and the fitting is by maximum likelihood estimation.

Fig. 4. For Adelaide Airport, comparisons of runs of upper extremes (90 per cent thresholds) for (a) Wind run, (b) evaporation, and lower extremes (10 per cent thresholds) for (c) sunshine, and (d) mean sea level pressure at 9.00 am. Observed (filled circles), untuned model (lines).



Generality

In principle, the model applies to any daily variable that has daily persistence and is seasonal. Further, by invoking symmetry it is clear that runs of days below percentile thresholds may be characterised in a similar way. Thus for a run of low extremes, we would use ceiling thresholds of the 25, 24, 23,…2, 1st percentiles instead of floor thresholds of the 75, 76, 77,… 98, 99th percentiles.

Model performance

Several graphical examples of runs of extreme days for a range of meteorological variables are presented to provide a qualitative assessment of the model, before a quantitative assessment based on runs of extremely hot days and cold nights for Australian and European sites.

Adelaide: wind run, evaporation, sunshine and pressure

For Adelaide Airport (at Fig. 4), the untuned model is compared to observed runs for upper extremes (above the 90th percentiles) for (a) wind run, (b) evaporation, and lower extremes (below 10th percentiles) for (c) sunshine, and (d) mean sea level pressure at 9.00 am. Subjectively, Fig. 4 shows good agreement between model and observations. Also shown are the absolute values corresponding to the given percentiles. Although not presented here, corresponding graphical comparisons for other capital city airport sites appear to have similar level of agreement between observations and model.

Some Australian sites using temperature time series up to 1970

Using the available daily maximum and minimum temperatures up to 1970, eight Australian sites (Alice Springs and seven capital cities) were analysed. Canberra and Brisbane are shown although they are exceptions to the requirement for at least 25 years of complete data. Model and observed frequency of runs for daily maximum temperatures exceeding the 90th percentiles (informally described here as heatwaves) and likewise for daily minimum temperatures below the 10th percentiles (informally described here as cold spells) are shown graphically at Figs 5 and 6. The model estimates use the a priori values of a and s. Several of the plots show very good agreement although the comparisons at Canberra for heatwaves and Alice Springs for cold spells are poorer. Also shown are model estimates using the tuned values of a^* and s^* which are shown in brackets. A performance score, *RMSEA*, which is described below, is also shown together with a value for the tuned model in brackets. The fine lines represent the untuned model and the bold line represents the tuned model, so it can be seen that the comparisons are improved for some sites. Confidence intervals (of 95 per cent) derived from Monte Carlo simulations (10 000 simulations of the model) are plotted to provide an appreciation of the sampling variability expected from the model given the number of years of observational data. They also give a graphical impression as to how well the model fits the observational data. Plots (not shown) for all other Australian sites and for detrended time series up to 2010 are subjectively similar.

Some European sites with long temperature records up to 1970

At Fig. 7 are similar comparison plots for eight European sites, for heatwaves based on daily maximum temperatures above the 90th percentiles. These sites were selected because they had the most years of available record. As with the Australian sites, the plots show good model fits with sometimes only marginal improvement between the untuned and tuned model. Plots (not shown) for runs based on daily minimum temperatures below the 10th percentiles are similar. These plots cover up to 200 years of record and indicate that the observations agree reasonably with the model in the tail area out to runs lengths of 15 or more days.

Although it is impractical to present the corresponding graphs for all Australian and European sites in the datasets, it is remarked that the comparisons between observations and model are subjectively very similar to those of the previous Figs. This comment applies to other thresholds such as 95th, 98th and 5th and 2nd percentiles.

Quantitative assessment

Quantitative assessment of the model's accuracy and reliability is performed for each of the temperature datasets for heatwaves (cold spells), for thresholds of 90, 95 and 98 per cent (ten, five and two per cent) for the untuned model and the tuned model. Performance measures were the correlation coefficient r, the χ^2 test, and RMSEA.

The χ^2 goodness of fit test (following Conover 1999) gives a binary outcome of 'acceptance or rejection' at the 0.05 significance level for the null hypothesis, in this case, that the observed number of runs has the distribution described by Equation 3. The percentage of sites for which the model is accepted (strictly, 'not rejected') is the performance measure and is denoted F. A disadvantage of this measure is that the χ^2 goodness of fit test tends to over-reject with large sample sizes, more than a few hundred (Kline 2011), as is the case here. For example, for the Australian cities with 60 years of record, the sample size ~ 1200 for thresholds of 90 per cent.

RMSEA (root mean square error of approximation) is a measure popular in the structural equation modelling community that is based on the χ^2 value but is adjusted for sample size N (Kline 2011). A value up to 0.05 is regarded as indicating a 'close approximate fit', a value between 0.05 and 0.08 indicates a 'reasonable approximate fit' and more than 0.1 indicates a poor fit. RMSEA is defined in Equation 5 (adapted from Kline, 2011) as

$$RMSEA = \sqrt{\frac{\max\left(\frac{\chi^2}{\nu} - 1, 0\right)}{N-1}} \qquad ...(5)$$

where ν is the number of degrees of freedom. The tuned model has two fewer degrees of freedom on account of the

Fig. 5. Frequency of runs of days hotter than the 90th percentile, informally referred to here as heat waves. Observed (filled circles), untuned model (fine lines), tuned model (solid lines), for eight Australian sites with data up to 1970. The use of the 90th percentile as a threshold is arbitrary and the model is applicable to other percentiles. Monte Carlo-simulated confidence intervals (CI) are represented by the vertical bars: fine lines offset slightly left for the untuned model and thick lines offset right for the tuned model. *CI* are for 95 per cent (lower and upper limits of 2.5 and 97.5 per cent respectively). T_x refers to maximum temperature.

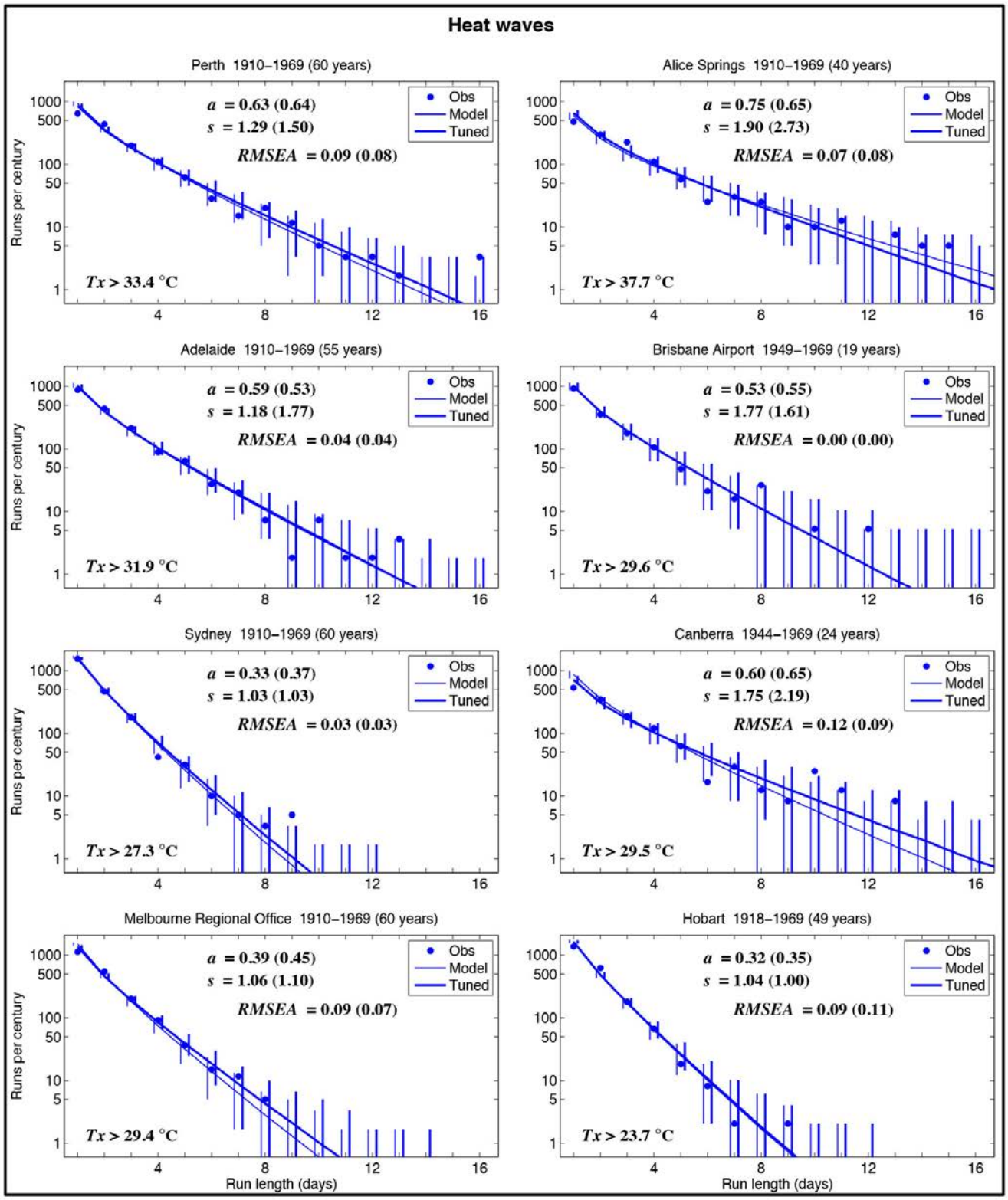


Fig. 6. Similar to Fig. 5 but for frequency of cold spells based on minimum temperatures below the 10th percentile. T_n refers to minimum temperature.

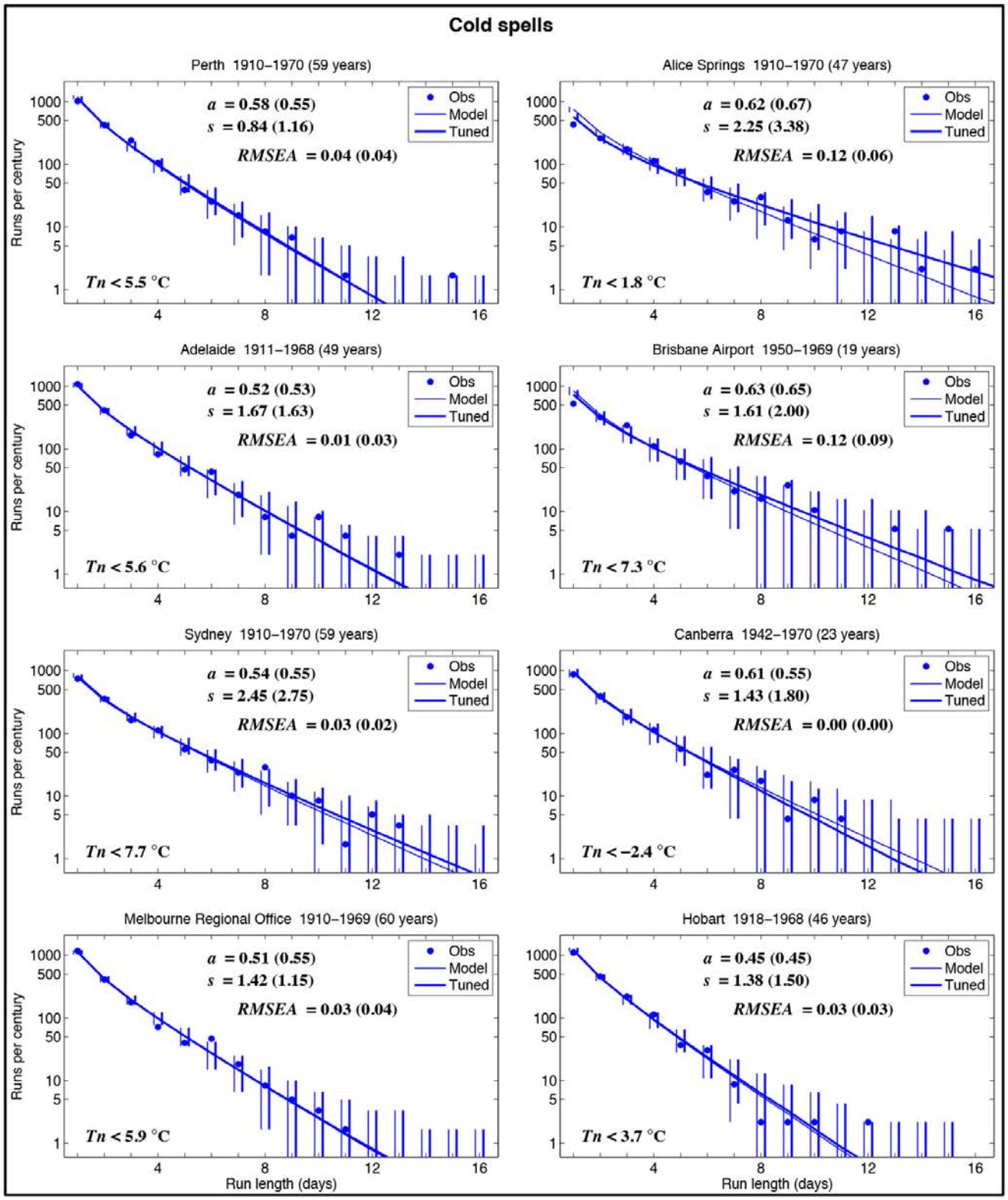


Fig. 7. Similar to Fig. 6 but for eight European sites with long records (100~200 years).

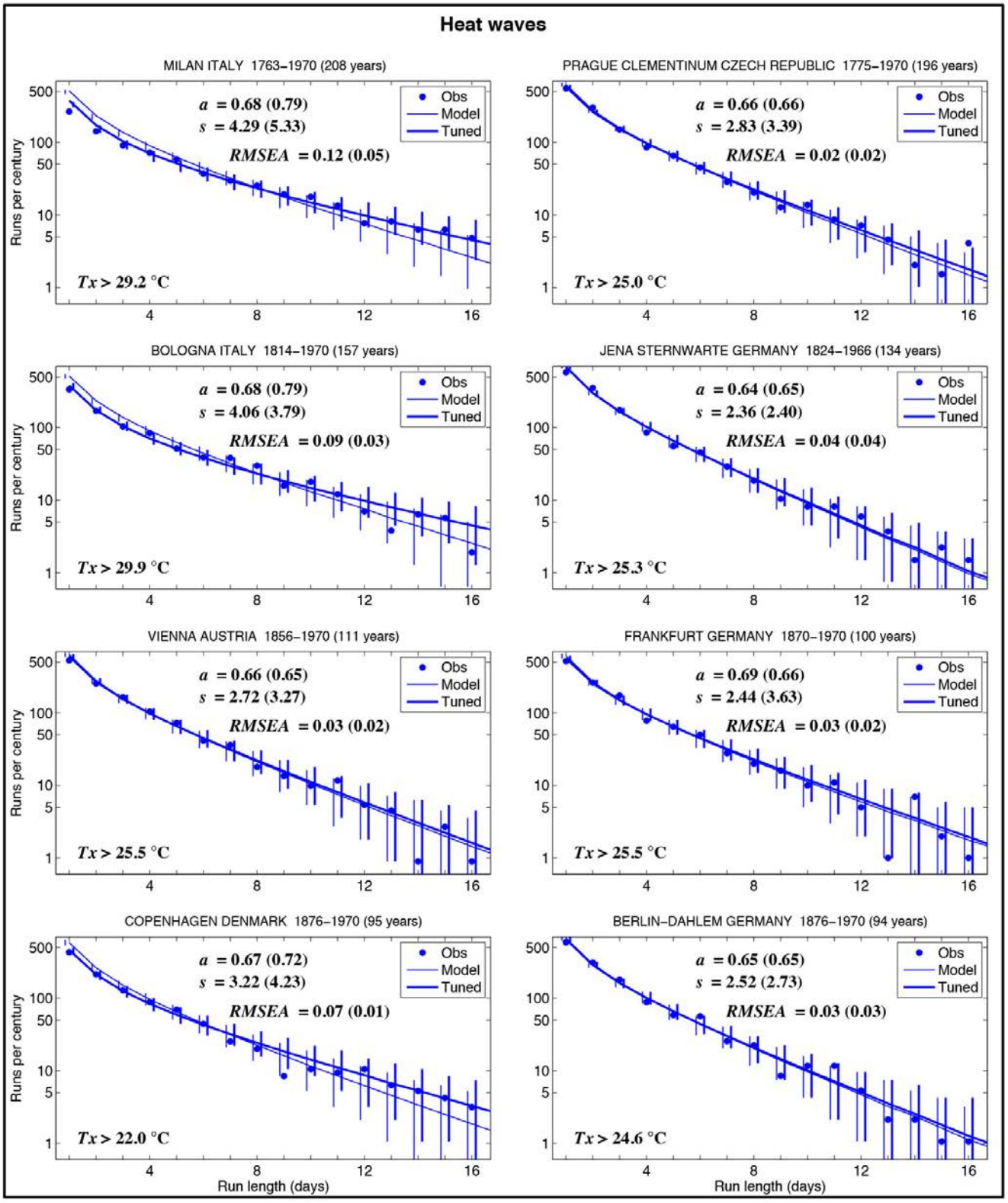
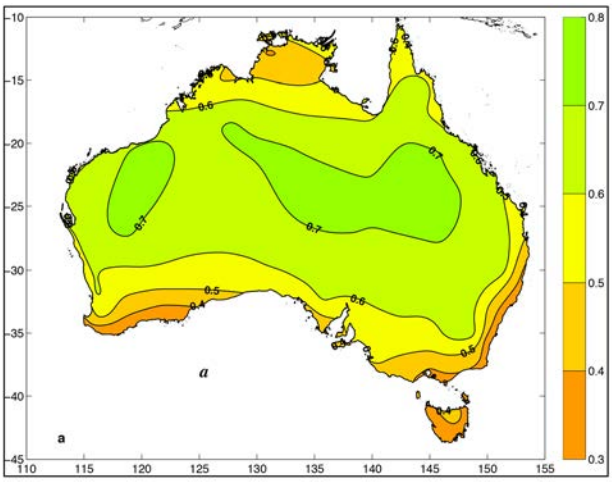
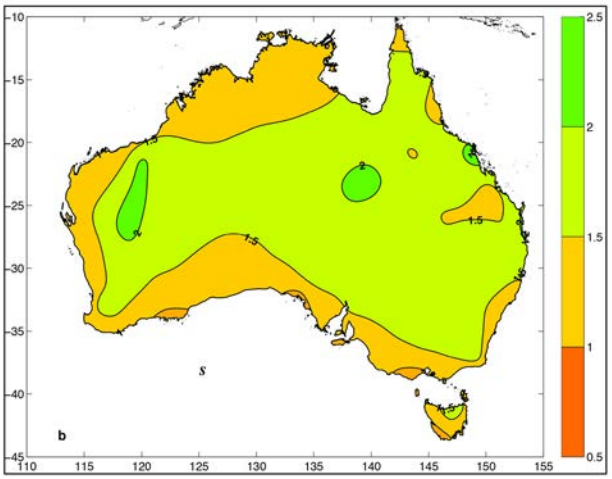


Fig. 8. Contours of (a) autocorrelation a and (b) seasonality index s for heat waves as calculated for Australian sites. Each of autocorrelation and seasonality are maximum in central Australia and minimum about southern coasts.

a)



b)



two fitted parameters. The term χ^2/ν is called the normed chi-square and a convenient rule of thumb suggested by some authors is that if it is less than two or three then the fit is acceptable, but Kline advises against this practice.

For each of the sites, with very few exceptions, the correlation coefficients between observed and modelled frequencies are about 0.97 or more. This tendency to be apparently over-generous arises because the correlation process favours a straight line fitted through the two largest values—the other values are one or more orders of magnitude smaller and therefore contribute little to the error terms. Even using transformed variables (square root values or logged values provided the observation counts are at least one) results in correlation coefficients that are rarely below 0.9. The correlation coefficient for the logged values of model and observed runs is used here: at about 0.95 they are quite high as reported at Table 1.

The overall F values are 22 per cent and 61 per cent for untuned and tuned model respectively implying that the χ^2 test rejects the untuned model about 80 per cent of the time while the tuned model is accepted about 60 per cent of the time. The overall mean *RMSEA* values for the untuned and tuned model respectively are 0.08 and 0.05. Using Kline's suggested *RMSEA* criteria above, then the untuned model is a reasonable approximation while the tuned model is usually a close approximation. The tuned model outperforms the untuned model. Performance between datasets is broadly comparable: however, the performance for heatwaves is better than that for cold spells, and for cold spells the performance with the European dataset is slightly worse than that of the Australian dataset. Bearing in mind r, F, and *RMSEA* then it is concluded that the untuned model is typically fair while the tuned model is typically fair to good.

A consistency check was performed on the spatial variability of the parameters a and s on the assumption that for physically meaningful parameters there should be spatial coherency. Contour plots of the parameters a and s for heatwaves are provided at Fig. 8. These show that each of a and s tend to increase inland or with greater continentality and a tendency for the lowest values to occur on the southern coasts. Tryhorn and Risbey (2006) used gridpoint data from NCEP/NCAR Reanalysis and GCM (global climate model) projections in their study on the distribution of heatwaves over Australia. They defined a heat wave as a run of days with maximum temperature exceeding the 90th percentile. They found longer heatwaves in central Australia and shorter heatwaves along the south coast of Australia which they attributed to the relatively static character of the central Australian weather and to the high-frequency frontal nature of weather in the southern regions respectively. This is consistent with the expectation from the model for higher a and s in central Australia and lower values about the southern coats. Using Fig. 3, as a and s increase we move right and down and see that the short runs become fewer and the longer runs become more common; as a and s decrease, we move toward the upper left where the short runs are more frequent and the long runs are rarer. A practical application for contour plots like Fig. 8 might be to interpolate to a site with little or no data.

Discussion and conclusions

The model is based on the idea that, in a descriptive statistical sense, the two most important causes of runs of extreme days for any meteorological variable are the autocorrelation during the peak season for extremes and the seasonality. The importance of autocorrelation is self-evident: the importance of seasonality is readily apparent when it is considered that regardless of autocorrelation, if the 'quota' of extreme days is confined to a short peak season then there is more chance of longer runs of extremes than otherwise. This is in part because the threshold for extremity is defined with respect to the whole year. Over and above the assumption of first

order autoregression, the presented model disregards the following factors or complications:

(a) The mean only approximately follows a sine curve.
(b) The peak periods for the extremes do not necessarily occur in January or July.
(c) The standard deviation is not constant through the year.
(d) The autocorrelation is not constant through the year.
(e) Higher order autoregression may be more appropriate.
(f) The stochastic shock element is drawn from a normal distribution. At least for summer-time maximum temperatures for southern Australia the observed distribution is better fit by a bi-normal distribution (Grace and Curran 1993, Trewin 2001).
(g) The seasonality index is a ratio of two independent component measures and therefore is more prone to error (10 per cent errors in each of the components could result in 20 per cent error in the ratio).
(h) Clustering of extremes could occur due to larger scale influences such as ENSO.
(i) Local or regional scale feedbacks might temporarily alter the autocorrelation. An example might be the grass and soil drying effect from a run of hot dry days which then preconditions a subsequent day to be hotter than it would have been otherwise. Similarly for winter time cold spells in the European locations, snow cover or ice formation associated with a run of extremely cold days could increase the chances of the subsequent day(s) being extremely cold.
(j) The values of autocorrelation and seasonality were assumed constant in the long term but could vary under climate change. For example, the autocorrelation of summer-time temperatures would rise if anticyclonic blocking increased or if summer-time cold fronts became less prevalent.

For the untuned model, these complicating factors are ignored or assumed self-cancelling; for the tuned model, it is assumed that all or most of the complications can be accommodated by adjustments to a and s to form effective parameters, namely, effective autocorrelation a^* and effective seasonality s^*.

Applying the model to the particular variable of maximum temperatures in Australia, it was shown that the untuned parameters have a spatial coherence. The model implication of the spatial pattern of the a and s parameters is for longer duration heatwaves in central Australia and shorter duration heatwaves about southern coastal regions which is as observed by Tryhorn and Risbey (2006) for both historical and GCM projections.

A set of graphical comparisons of the untuned model and observed runs frequency of extreme days of high wind run, high evaporation, low sunshine hours and low pressure at Adelaide showed good qualitative agreement. For the heatwaves and cold spells at the eight Australian sites the model, when tuned, showed very good agreement. This was also true for a set of European sites, selected as those eight with the longest available record—about 100 to 200 years.

Complications due to global warming trends were avoided by restricting the quantitative testing of the model to the period up to 1970, however, it is suggested that appropriate detrending of the relevant time series would allow the model to be applied to quantities exhibiting a secular trend. If this were done then the percentile thresholds of the detrended time series would be used and the modelled frequency of runs would be applicable to any era of the time series. For the untuned model, the values of a and s are initially calculated yearly and then meaned over the period of record. As noted earlier, these yearly values are not affected by a secular trend in the variable concerned. This opens the possibility of investigating a and s for trends. If they were found to be increasing then the implication would be that longer runs of extremes of the variable would become more frequent at the expense of shorter runs (see Fig. 3), regardless of any trend in the variable itself.

Quantitative tests of the model used approximately 70 Australian sites from a high quality daily maximum and minimum temperature dataset and approximately 60 European sites—many with records of 100 years or more. For each site, separate tests were performed for runs of extreme days, hot and cold, for thresholds of 90, 95 or 98 per cent and ten, five or two per cent respectively. Judged on performance measures of correlation coefficient, the percentage accepted by the χ^2 goodness of fit test, and *RMSEA*, the untuned model typically gave fair agreement with the observations, while the tuned model typically gave fair to good agreement. It is concluded that the model has the potential to estimate frequency (or return period) of unusually long runs—those with duration at least up to ~15 days. In principle, the meteorological quantities of interest are general and may potentially include many meteorologically related quantities such as lake level, pollen counts and fire danger indices.

Acknowledgments

Reviewer and editor comments lead to much greater readability. This paper was developed from previous research funded by the Grape and Wine Research and Development Corporation.

References

Arnau, J. and Bono, R. 2002. A program to calculate the empirical bias in autocorrelation estimators. *Psicothema, 14*, 669–72.

Bureau of Meteorology. 2011. Data available at www.bom.gov.au/climate/data/. Accessed 15 August 2011.

Bureau of Meteorology. 2012. Data available at www.bom.gov.au/climate/change/acorn-sat/. Accessed 1 May 2012.

Commonwealth of Australia. 2012. *State of the Climate 2012*. A joint publication of Bureau of Meteorology and CSIRO. pp12.

Conover, W. J. 1999. *Practical Nonparametric Statistics. 3rd ed.* Wiley. pp584.

Grace, W. and Curran, E. 1993. A binormal model of frequency distributions of daily maximum temperature. *Aust. Meteorol. Mag., 42,* 151–61.

Grace, W.J., Sadras, V.O. and Hayman, P.T. 2009. Modelling heatwaves on viticultural regions of southeastern Australia. *Aust. Met. Oceanogr. J., 58,* 249–62.

Grace, W.J. 2011. Modelling heatwaves: Connecting an Empirical Markov Model with an Autoregressive Model. *Aust. Met Oceanogr. J., 61,* 43–52.

Kirchner, J. 2001. Analysis of Environmental Data: Data analysis toolkit 11. University of California, Berkeley. seismo.berkeley.edu/~kirchner/eps_120/EPSToolkits.htm. Accessed 1 June 2012.

Klein Tank, A.M.G., J.B. Wijngaard, G.P. Können, R. Böhm, G. Demarée, A. Gocheva, M. Mileta, S. Pashiardis, L. Hejkrlik, C. Kern-Hansen, R. Heino, P. Bessemoulin, G. Müller-Westermeier, M. Tzanakou, S. Szalai, T. Pálsdóttir, D. Fitzgerald, S. Rubin, M. Capaldo, M. Maugeri, A. Leitass, A. Bukantis, R. Aberfeld, A.F.V. van Engelen, E. Forland, M. Mietus, F. Coelho, C. Mares, V. Razuvaev, E. Nieplova, T. Cegnar, J. Antonio López, B. Dahlström, A. Moberg, W. Kirchhofer, A. Ceylan, O. Pachaliuk, L.V. Alexander, and P. Petrovic. 2002. Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. *Int. J. Climatol., 22,* 1441–53. (Data and metadata available at http://eca.knmi.nl). Accessed 30 Dec 2011.

Kline, R. B. 2011. *Principles and practice of structural equation modeling*. 3rd ed. The Guilford Press, New York. Pp 427.

Klok, E.J, and Klein Tank, A.M.G. 2009. Updated and extended European dataset of daily climate observations. *Int J. Climatol., 29,* 1182–91.

Kysely, J., 2010. Recent severe heatwaves in central Europe: how to view them in a long-term perspective? *Int J. Climatol., 30*, 89–109.

Law, A.M. and Kelton. D.W. 1991. *Simulation Modeling and Analysis*. 2nd Ed. McGraw-Hill, p284.

Mearns, L.O., Katz, R.W., and Schneider S.H. 1984. Extreme high temperature events: changes in their probabilities with changes in mean temperature. *J. Clim. Appl. Meteorol., 23,* 1601–08.

Perkins, S. E., and Alexander, L. V. 2013. On the Measurement of Heatwaves. *J. Climate, 26,* 4500–517. doi: dx.doi.org/10.1175/JCLI-D-12-00383.1

Seaman, R.S. 1992. Serial correlation considerations when assessing differences in prediction skill. *Aust. Meteorol. Mag., 40,* 227–37.

Trewin, B.C. 2001. *Extreme temperature events in Australia*. PhD thesis, School of Earth sciences, University of Melbourne, Australia.

Trewin, B. 2013. A daily homogenized temperature data set for Australia. *Int. J. Climatol., 33,* 1510–29. doi: 10.1002/joc.3530

Tryhorn, L. and Risbey, J, 2006. On the distribution of heatwaves over the Australian region. *Aust. Meteorol. Mag., 55,* 169–182.

Von Storch, H. and Zwiers, F.W. 1999. *Statistical Analysis in Climate Research*. University Press, Cambridge. 484pp.

Wilks, D.S. 2008. High-resolution spatial interpolation of weather generator parameters using local weighted regression. *Agric. Forest Meteorol., 148,* 111–20.

Wilks, D.S. 2011. *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Academic Press. 676 pp.

Appendix A

The distribution of runs of extreme days is shown to be of an exponential nature in the special case when the meteorological quantity concerned is without autocorrelation and seasonality. In such a case its time series may be represented as a random variable. Let the threshold of an upper extreme be taken as the p^{th} percentile quantity from a very long stationary sequence.

The fraction f of days when the quantity exceeds p^{th} percentile is $f = 1 - 0.01p$, and where $0 < f < 1$ in practice. Now f is independent of all previous days. Thus f is the probability that a given day is extreme and $1 - f$ is the probability that a given day is not extreme.

In the example at Fig. 1, the probability P of a run of three days starting from any arbitrary day is given by the product $(1 - f) f^3 (1 - f)$ arising from the constraint that a run of three extreme days must be preceded by a non-extreme day and succeeded by a non-extreme day. Generalising to a run of x days ($x = 1,2,3\ldots$), then

$$P(x) = (1 - f)^2 f^x. \qquad \ldots(A1)$$

The probability function P is easily shown to be exponential in x. Taking logarithms, rearranging and exponentiating Equation A1 gives

$$P(x) = (1 - f)^2 \exp\left[(\ln f)x\right]. \qquad \ldots(A2)$$

Casting Equation A2 into expected runs in a period of a century then

$$R(x) = (365)(100)(1 - f)^2 \exp\left[(\ln f)x\right]. \qquad \ldots(A3)$$

Replacing f with $1 - 0.01p$, then Equation A3 becomes

$$R(x, p) = 3.65\, p^2 \exp\left[\left(\ln(1 - 0.01p)\right)x\right] \qquad \ldots(A4)$$

where R is the number of occurrences per century for runs of length $x = 1, 2, 3,\ldots$ days when the variable concerned exceeds percentile thresholds of p as in the main text. Since $p < 100$, then $\ln(1 - 0.01p)$ is negative and thus R reduces exponentially with x. The simulation plots of the top-left panel in Fig. 3 are reproduced by Equation A4.

Appendix B

It is known that the sample autocorrelation and sample standard deviation of an autocorrelated variable is biased (Law and Kelton 1991, Seaman 1992, Kirchner 2001, Arnau and Bono 2002). Generally for positive autocorrelation the bias is negative (under-estimation occurs). Kirchner (2001) suggests the following viewpoint. Since the data are not independent then the residuals from a mean or any model are largely redundant (i.e., not independent of one another). Therefore the effective degrees of freedom are far fewer than the number of observations. Then for those statistical parameters (other than the mean) where the sample size appears as a denominator, the sample estimate of the parameter concerned will be an under-estimate of the population parameter.

Arnau and Bono (2002) provide a Monte Carlo simulation method to estimate the bias of the sample autocorrelation as a function of the sample autocorrelation and sample size. The method is easily modified to provide the bias for the sample standard deviation. Following their method, the results for sample sizes of 30 and 61 are shown at Fig. B1. From the curves of sample size 30, empirical rational linear expressions for the bias corrections were easily obtained.

Fig. B1. Plots of bias correction against sample autocorrelation for (a) autocorrelation and (b) standard deviation derived from Monte Carlo simulation for sample sizes of 30 and 61.