Water Management
Volume 166 Issue WM10

A rainfall model for drought risk analysis in south-east UK
Duan, McIntyre and Onof

ice | proceedings

Proceedings of the Institution of Civil Engineers
Water Management 166 November 2013 Issue WM10
Pages 519–535 http://dx.doi.org/10.1680/wama.12.00012
Paper 1200012
Received 11/01/2012 Accepted 16/07/2012
Published online 22/02/2013
Keywords: hydrology and water resource/risk and probability analysis/water supply

ice Institution of Civil Engineers publishing

A rainfall model for drought risk analysis in south-east UK

Juan Duan BEng, MSc
Postgraduate researcher, Department of Civil and Environmental Engineering, and Grantham Institute for Climate Change, Imperial College London, UK

Neil McIntyre BEng, MSc, PhD, MICE, CEng
Reader in surface water hydrology, Department of Civil and Environmental Engineering, Imperial College London, UK

Christian Onof PhD, FSS
Reader in stochastic environmental systems, Department of Civil and Environmental Engineering, Imperial College London, UK

Drought risk assessment ideally requires long-term rainfall records especially where inter-annual droughts are of potential concern, and spatially consistent estimates of rainfall to support regional and inter-regional scale assessments. This paper addresses these challenges by developing a spatially consistent stochastic model of monthly rainfall for south-east UK. Conditioned on 50 gauged sites, the model infills the historic record from 1855–2011 in both space and time, and extends the record by synthesising droughts which are consistent with the observed rainfall statistics. The long record length allows more insight into the variability of rainfall and potentially a stronger basis for risk assessment than is generally possible. It is shown that, although localised biases exist in both space and time, the model results are generally consistent with the observed record including for a range of inter-annual droughts and spatial statistics. Simulations show that some of the most severe inter-annual droughts on the record may recur, despite a trend towards generally wetter winters.

Notation

c	correlation coefficient
C_{ii}	variance of the unconditional error at site i
$\boldsymbol{C}_{si}$	vector of covariance values between s sites and another site i
$\mathbf{C}_{ss}$	covariance matrix describing the dependencies between the errors at s sites
D	distance between two sites
$\boldsymbol{E}$	$n \times 1$ vector of errors
$\bar{e}_i$	expected value of the error at site i
$\boldsymbol{e}_s$	vector of errors observed at s sites over all the years for any month
$N(\bar{e}_i, \sigma_i^2)$	a random sample drawn from a normal distribution with mean $\bar{e}_i$ and variance σ_i^2
r	untransformed rainfall
$\mathbf{X}$	$n \times m$ matrix containing n values of m observed input variables
$\boldsymbol{Y}$	$n \times 1$ vector of observations
y	transformed rainfall
$\bar{y}_i$	expected value of transformed rainfall
α	parameter of the correlogram
β	parameter of the correlogram.
θ	$m \times 1$ vector of regression coefficients
λ	Box–Cox transform parameter
σ_i^2	variance of the conditional error at site i

1. Introduction

The sustainability of water supply in much of Europe is a major concern for economic and environmental planning (Mechler and Kundzewicz, 2010). One region of particular concern is south-east UK, which has a high and increasing population, relatively low rainfall and high evaporation (Arnell and Delaney, 2006; Marsh *et al.*, 2007). A large proportion of the supply in this region is from the chalk aquifer, which is under stress in places from over-abstraction and agricultural contamination (Smith *et al.*, 2010). To relieve the stress on water resources, options for desalination, bulk imports and inter-basin transfers from the Thames and Severn basins have been considered (Arnell and Delaney, 2006). Moreover, towards more optimal sharing of water resources during drought periods, there are currently efforts to optimise water transfer schemes within the south-east (von Lany *et al.*, 2008).

In south-east UK it is generally perceived that three dry winters in succession would present severe regional water supply deficits (the winter season, with its higher rainfall and lower potential evaporation, being the primary source of effective rainfall and recharge to the aquifers) (von Lany *et al.*, 2008), and should the most extreme historic droughts recur (see Marsh *et al.*, 2007) it seems unlikely that an acceptable level of service could be maintained (McIntyre *et al.*, 2003). Of particular concern to water managers is the possible recurrence of the long-term droughts of 1887–1910, which included a series of five unusually dry winters, and shorter inter-annual droughts of 1920–1922, 1933–1934, 1975–1976, 1990–1992 and 1995–1997 (Marsh, 1996; Marsh *et al.*, 2007; Subak, 2000).

As well as drought duration, the spatial aspect of drought is also of interest (Burke and Brown, 2010; Zaidman *et al.*, 2002). The spatial properties of drought have particular practical relevance in

south-east UK, where extending the water network within and beyond the region is potentially viable (von Lany *et al.*, 2008). Assessing the scope for such transfers requires a good understanding of the spatial characteristics of water availability over the relevant scales. Hence there is considerable motivation for developing data sets and tools which deliver a capability for characterising both the temporal and spatial characteristics of extreme droughts.

Gridded climate re-analysis data sets (e.g. the ERA40 data, Uppala *et al.* (2005)) produce historical rainfall going back to 1957 on a grid scale of about 1·125°. However this time coverage and grid scale are rather restrictive for analysis of extremes within one region. Downscaling tools provide a simulation capability at more applicable scales (e.g. Kenabatho *et al.*, 2011; Kigobe *et al.*, 2011). For example, the weather generator associated with UKCP09 (Jones *et al.*, 2009) can be used to generate scenarios of extreme rainfall on a 5 km grid scale. However, that weather generator was not designed to produce spatially consistent rainfall in the sense that observed inter-grid dependence of rainfall is not preserved (Jones *et al.*, 2009), and hence has limitations for spatial assessment of drought risk. Also, many stochastic rainfall models (e.g. Jones *et al.*, 2009; Yang *et al.*, 2005) are trained on only around 30 years of data, and hence their suitability for generating a range of extreme droughts is unclear. Therefore, despite their attractions, existing re-analysis data sets and stochastic rainfall models are not by themselves adequate to support regional drought risk analysis.

The concern about water stress in England, Europe and beyond calls for suitable data sets and tools to support regional water resources management (Thyer *et al.*, 2002). This includes generating long sequences of spatially distributed rainfall over space and time. This paper aims to address this challenge by developing new statistical rainfall models using a case study of south-east UK, including the following activities.

(*a*) Compilation of available long-term rainfall records covering south-east UK (Kent, Sussex, Hampshire, Surrey, Isle of Wight, east Wiltshire, south Berkshire and south London).
(*b*) Identification of large-scale climatic drivers of rainfall and regional variability to give a deterministic model to predict expected rainfall over the region; and identification of a stochastic model to describe variability around the expected values.
(*c*) Infilling of missing data to provide continuous monthly sequences of rainfall dating back to 1855, gridded over the south-east region, including uncertainty estimates.
(*d*) Assessment of the ability of the model to replicate the historical extreme droughts, in particular the severe droughts of 1887–1910, 1920–1922 and 1975–1976.

2. Rainfall analysis using regression

Statistical modelling has commonly been employed to infill partial historical rainfall sequences and to downscale climate model projections (Fowler *et al.*, 2007; Xu, 1999): regression is one such approach (Hanssen-Bauer and Førland, 1998; Murphy and Washington, 2001; Phillips and McGregor, 2002). Multiple linear regression may be described by

$$\boldsymbol{Y} = \mathbf{X}^{\mathrm{T}}\theta + \boldsymbol{e} \quad (1)$$

where $\boldsymbol{Y}$ is a $n \times 1$ vector of observations, $\mathbf{X}$ is a $n \times m$ matrix containing n values of m observed input variables, θ is a $m \times 1$ vector of constant regression coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of errors. θ is generally estimated by minimising the sum of the squared errors given the set of observations, $\boldsymbol{Y}$ and $\mathbf{X}$. With the assumption that the errors in vector $\boldsymbol{e}$ are independent of each other, and are identically and normally distributed, the least squares estimate of θ is equivalent to the maximum likelihood estimate. This assumption also allows the covariance matrix of the regression coefficients to be estimated using standard linear methods (Kottegoda and Rosso, 2008). The input variables to include in $\mathbf{X}$ are generally identified by trial and error, aiming to produce a model which explains much of the variability in $\boldsymbol{Y}$ (generally measured using the R^2 statistic), and also, ideally, to produce a θ with low covariance. Stepwise regression (Draper and Smith, 1998) is a set of procedures which assists with the identification of the optimal $\mathbf{X}$ variables (from a set of pre-specified candidates).

The identification of a suitable probability density function to describe $\boldsymbol{e}$ means that Equation 1 may be employed as a stochastic model, from which random realisations of $\boldsymbol{Y}$ can be simulated. This potentially provides a model for stochastic simulation of rainfall variability and extremes. Consistent with the general statistical assumptions behind least squares regression, it is common to assume a normal distribution of errors. Towards achieving such a normal distribution, the skewness generally observed in rainfall data can be managed by transforming the rainfall prior to the regression, for example using a logarithmic or Box–Cox transform (Kottegoda and Rosso, 2008). When the errors are not independent of each other (as in the case study below), a multivariate normal distribution is required. Where the rainfall sample contains a significant number of zeros, as would be the case using daily or sub-daily data in the UK, the random variability cannot conveniently be described by a single continuous distribution function. Furthermore, at these time scales there is significant serial dependence. These challenges have led to the use of statistical methods for rainfall modelling which are more flexible than simple regression (Chandler and Wheater, 2002, Segond *et al.*, 2006). However, in this paper, the use of *monthly* rainfall data sufficiently simplifies the problem so that a stochastic model of the form of Equation 1 (including suitable Box–Cox transforms of the data and suitable models of inter-site dependence) is proposed as sufficient.

3. A monthly rainfall model for south-east UK

3.1 Definition of 'south-east UK'

The 'south-east UK' is defined here as the region illustrated in Figure 1, bounded to the south and east by the coast, to the west by (using the UK national grid coordinate system) easting 410000 m and to the north by 180000 m. This spatial covergage was governed by: (1) the wish to cover a large part of south and south-east UK; (2) the increased difficulty of achieving a satisfactory spatial model if extending the region further north and/or west; and (3) the computational demands of stochastic modelling, which inhibit the inclusion of many more sites. Therefore, no particular climatic, geographical, political or water company boundaries were used to define the coverage, and they would need to be reviewed prior to a practical application of the model at the regional scale.

3.2 The climate of south-east UK

South-west frontal systems dominate the rainfall of south-east UK, hence rainfall generally reduces towards the east and north. As over the UK in general, significant correlations between rainfall and the North Atlantic Oscillation, and other variables and indices related to the Atlantic low pressure systems, are observed particularly in the winter months (Lavers *et al.*, 2010; Murphy and Washington, 2001; Wilby *et al.*, 2004; Yang *et al.*, 2005). Other climate indices reported to have some influence on rainfall in this region include the East Atlantic pattern (Barnston and Livezey, 1987) and storm track blocking indices (Pelly and Hoskins, 2003). The south-east is hotter and more humid in the summer than the rest of the UK, and convective type rainfall is significant. The average annual rainfall over the case study region is 730 mm, ranging from 524 mm in the dry north of Kent (site 6762 in Figure 1) to 982 mm in the relatively high altitude coastal South Downs (site 7504 in Figure 1). Considering regional-average annual rainfall (based on the infilled data set presented later in the paper), the standard deviation during the period 1855–2011 was 105 mm, the minimum was 430 mm (1921) and the maximum was 1017 mm (1960). The UKCP09 analysis (Jenkins *et al.*, 2008) did not find a significant trend (95% level) in either summer or winter rainfall in the south-east over the period 1914–2006 (their analysis included the whole of the Thames basin). Significant droughts in the south-east have included the long droughts from 1887 to 1910, and the shorter but more severe 1920–1922, 1975–1976 and 2004–2006 droughts (Marsh *et al.*, 2007).

3.3 Data sets

The rainfall data used in this study originate from the UK Meteorological Office MIDAS database. Details of the rain gauge network and recording practices can be found on the Hadley Centre website (see Table 1). Twenty-eight of the rain gauges provide long-term data (defined here as more than 80 years), and almost all gauges have considerable periods of missing data. In this study, the 28 long-term gauges were used to fit the rainfall model, supplemented by 22 shorter-term gauges to provide a spatially representative set. The gauge numbers and locations are shown in Figure 1, and the extent and continuity of data are shown in Figure 2. The data period used was from March 1855 (the earliest record available, at the Southampton East Park

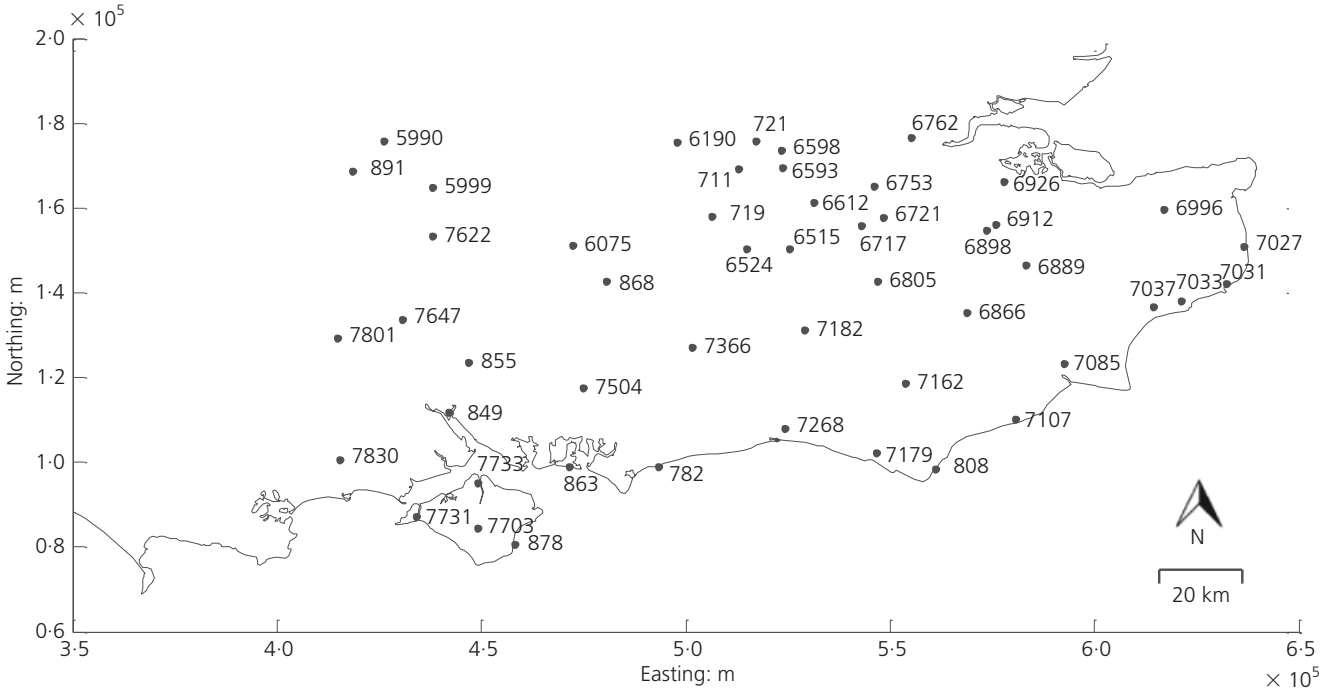

Figure 1. Gauge sites with outline of south-east UK

Data	Definition	Units	Period available	Data source	Website
Mean sea level pressure (MSLP)	The Met Office Hadley Centre's mean sea level pressure data set, HadSLP2, on a 5° latitude–longitude grid	mbar	1850–2004	Met Office Hadley Centre observations datasets	http://www.hadobs.org/
Central England temperature (CET)	Representative of a roughly triangular area of the UK enclosed by Bristol, Lancashire and London	°C	1659–2010	Met Office Hadley Centre observations datasets	http://www.hadobs.org/
North Atlantic oscillation (NAO)	Normalised pressure difference between Gibraltar and Reykjavik, Iceland	—	1821–2010	University of East Anglia Climatic Research Unit	http://www.cru.uea.ac.uk/cru/data/nao/
Atmospheric carbon dioxide	From 1958–2008, the Mauna Loa air intakes; from 1855–1957 from a spline of the Law Dome DE08 and DE08-2 ice cores	PPM	1832–1978/ 1958–2010	The Carbon Dioxide Information Analysis Center	http://cdiac.ornl.gov/
Trend	A linear trend	years	—	—	—
Elevation	Above UK Ordnance Datum (Newlyn)	m	—	British Atmospheric Data Centre	http://badc.nerc.ac.uk/
Northing	UK National Grid reference	m	—	British Atmospheric Data Centre	http://badc.nerc.ac.uk/
Easting	UK National Grid reference	m	—	British Atmospheric Data Centre	http://badc.nerc.ac.uk/

Table 1. Definitions and sources of predictor variables

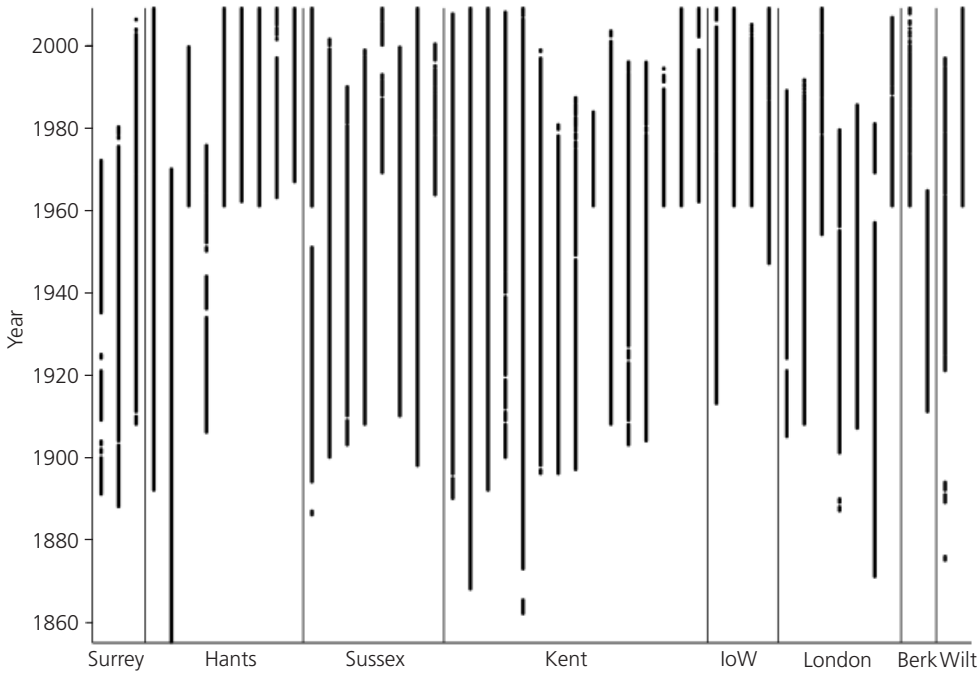

Figure 2. Record continuity at each gauge. The black bars indicate months with data

gauge) to December 2011. The daily data were aggregated to monthly; any month which contained one or more missing days was considered to be a missing month (to be infilled by the model). The monthly time-series were checked for inconsistencies and, for each gauge, any months with clearly perceived quality problems were removed (44 values of monthly rainfall in total).

Monthly climate data used as inputs to the model were selected according to the availability of long-term records and according to indications from the literature about their possible importance (Barry and Chorley, 2003; Hulme and Barrow, 1997). These climate variables are: the North Atlantic Oscillation index, Central England temperature, Mean Sea Level Pressure, and the East Atlantic index. Central England air temperature (as opposed to more local air temperature) is used because it spans the rainfall time period of 1855–2011 and at a monthly scale it is almost perfectly correlated with the south-east regional average temperature during the period 1914–2006 (correlation coefficient = 0·99). The spatial inputs are: northing and easting on the UK national grid coordinate system in units of metres, and altitude in units of metres above sea level. The definitions, origins and time periods covered by the data sets are listed in Table 1.

3.4 Deterministic component of the model

The aim of the regression is to identify a model which characterises the space and time variability of rainfall, and allows simulation. This includes a deterministic component (which estimates the expected rainfall given the input variable values for any month for any location) and a stochastic component (to estimate variability around the expected value including inter-site dependence). The analysis methods are essentially empirical, although any models found to be inconsistent with known physical relationships would be rejected. All modelling was done using Matlab version R2010b.

In the study presented here the general regression model in Equation 1 is applied where $\boldsymbol{Y}$ is the vector of rainfall observations including all 50 sites, and $\mathbf{X}$ is the corresponding values of the predictor variables in Table 1. In pursuit of a normal distribution of regression errors, a one-parameter Box–Cox transform (Kottegoda and Rosso, 2008, p. 366) is applied to the monthly rainfall data before the regression model is fitted

$$y = \frac{r^{\lambda} - 1}{\lambda} \quad (2)$$

where y is the transformed rainfall sample (i.e. a sample from the $\boldsymbol{Y}$ vector), r is the corresponding untransformed value (in mm/month) and λ is the Box–Cox parameter which is optimised to minimise skewness of the error distribution. After the model is applied, y is transformed back into r using the inverse of Equation 2. Each of the input variables in $\mathbf{X}$ was normalised so that its sample had zero mean and unit variance. This transformation allows the magnitudes of the optimised regression coefficients to be interpreted as relative sensitivity measures (Draper and Smith, 1998; Tabachnik and Fidell, 1996).

An independent regression model is developed for each of the 12 months. While this divides the data set into 12 and hence restricts the number of data points available per model, this month-by-month approach has the advantage that it allows the seasonal variability of the rainfall to be characterised by the model coefficients rather than imposing an approximate seasonal structure. Despite splitting of the data set into 12, the long-term data and multiple sites ensure that there are sufficient data to identify statistically significant models.

Within this regression framework, the model may be fitted either to a single site, where the vector $\boldsymbol{Y}$ contains transformed rainfall data from only one rain gauge and matrix $\mathbf{X}$ contains no spatial information, or to multiple sites, where $\boldsymbol{Y}$ contains data from multiple gauges and $\mathbf{X}$ contains spatial input variables which aim to explain the variation in expected rainfall between gauges. Only the multi-site analysis results are presented herein.

3.5 Stochastic component of the model

The deterministic regression of transformed rainfall allows identification and analysis of significant input variables, and infilling expected values of monthly rainfall at gauged and ungauged sites. However, to represent variability around the expected value a stochastic error model is also required. This allows the uncertainty in reconstructing partially observed events such as those in 1897–1910 to be modelled, and is essential for the simulation of possible but yet unobserved extreme droughts. The Box–Cox transform allows the errors to be approximately normally distributed with zero mean, hence the error model for any one month for a single site is straightforward. However, two types of error-to-error dependency potentially exist: dependency between errors from one month to another, and dependency between errors from one site to another. The former turns out to be insignificant (as confirmed in the results reported below); the inter-site dependency, as should be expected using monthly data from sites within one region, is crucial.

When infilling missing data at any one of the 50 gauged sites, the inter-site dependency of errors is treated in the following manner. The stochastic component of (Box–Cox transformed) rainfall at any site can be estimated conditional on the errors observed at the other sites using the standard procedure of generating samples from a multivariate normal distribution. This procedure is described in Searle (1971) and summarised here. Given a vector of errors observed at s sites over all the years for any month ($\boldsymbol{e}_s = \boldsymbol{Y} - \mathbf{X}^{\mathrm{T}}\theta$) and the covariance matrix describing the dependencies between the errors at these s sites ($\mathbf{C}_{ss}$) and the vector of covariance values between these s sites and another site i for which data are missing ($\boldsymbol{C}_{si}$), then the expected value of the error at site i is

$$3. \quad \bar{e}_i = \boldsymbol{C}_{si}^{\mathrm{T}}[\mathbf{C}_{ss}]^{-1}\boldsymbol{e}_s$$

Given an estimate of the variance of the unconditional error at the unobserved site (i.e. the variance of the error at site i irrespective of the other sites) (C_{ii}) the variance of the conditional error e_i is

$$4. \quad \sigma_i^2 = C_{ii} - \boldsymbol{C}_{si}^{\mathrm{T}}[\mathbf{C}_{ss}]^{-1}\boldsymbol{C}_{si}$$

For the set of 50 gauged sites, all of which have periods of overlapping data (Figure 2), the covariances can be estimated, so that $\boldsymbol{C}_{si}$, $\mathbf{C}_{ss}$ and C_{ii} are known for any s set of sites and any site i. A missing month of data at site i is then simulated as

$$5. \quad y_i = \bar{y}_i + N\left(\bar{e}_i, \sigma_i^2\right)$$

where $\bar{y}_i$ is the expected value from the deterministic component of the model and $N(\bar{e}_i, \sigma_i^2)$ signifies a random sample drawn from a normal distribution with mean $\bar{e}_i$ and variance σ_i^2. In principle, this method can be used to synthesise data for missing periods in the data record while approximating the observed spatial dependence structure. This would result (as far as the underlying model assumptions allow) in a spatially and temporally consistent historical time series. Furthermore, the stochastic nature of Equation 5 means that multiple realisations can be generated to represent the uncertainty associated with the infilling. For example, periods with few operating gauges will have relatively high uncertainty in regional rainfall, and sites at large distances from the nearest gauged sites will have relatively high uncertainty.

In practice, the direct use of the observed covariances in Equations 3 and 4 was problematic using the case study data, because $\mathbf{C}_{ss}$ was not positive-definite (Horn and Johnson, 1985), an indication that the sampled covariance is not consistent with a multivariate normal distribution. This is assumed to arise because the overlapping periods used to estimate $\mathbf{C}_{ss}$ were not the same for all pairs of sites, and so the sample used to calculate $\mathbf{C}_{ss}$ is not necessarily from a unique multivariate distribution. A potential solution is to form $\mathbf{C}_{ss}$ using only the sites nearest to site i. However, when tested, this only consistently resolved the problem when data from less than five sites were included, which is unlikely to produce an acceptable level of spatial consistency over the region. Instead, the problem of obtaining a real solution to Equations 3 and 4 was resolved by smoothing out the unwanted variability within $\mathbf{C}_{ss}$ by fitting a model of inter-site covariance, specified below, rather than directly using the sampled observations.

A model of inter-site covariance is obtained by identifying a correlogram model, where correlation between each pair of sites c is estimated as a continuous function of the distance between the two sites D. After testing various models, the following two-parameter equation was preferred

6. $c = \exp\left(-\alpha D^{\beta}\right)$

Also considering the difference in elevation between pairs of sites did not significantly improve upon this model. Parameters α and β were optimised using non-linear least squares using the observed inter-site correlations. Only pairs of sites with more than 50 years of overlapping data were used in this optimisation to reduce influence of less precise estimates of correlation. For any two gauged sites, multiplying the correlation by the observed standard deviation of errors at both sites gives an estimate of the covariance. Hence a smoothed version of the observed $\mathbf{C}_{ss}$ is obtained, which leads to a consistently real solution to Equations 3 and 4. The significance of using a modelled instead of observed inter-site error covariance will be tested as part of model verification.

The error model described above can be modified to allow extension of the historical record in space and time. Extension only in time requires generation of sets of errors over the 50 sites for months when no rainfall observations exist. Extension only in space means generating rainfall within the record period for hypothetical sites, for example to produce gridded rainfall. In this case (because i represents an ungauged site) rather than using an observed error variance in C_{ii} and C_{si}, a model is needed. This is approached by assessing whether and how error variance changes across the 50 gauged sites, and interpolating to the synthesised sites. Extending the record in both space and time combines these two modifications.

3.6 Model verification

The aim of model verification is, first, to test to what degree the statistical properties of the errors conform to the assumptions which have been made in model estimation. The specific tests carried out are listed here.

(*a*) Bias in errors over space and time.
(*b*) Deviation of errors from a normal distribution.
(*c*) Dependence of errors on input variables.
(*d*) Stationarity of variance in errors over space and time.
(*e*) Autocorrelation of errors between months.

Recognising that the properties of the errors will not exactly conform to the assumptions (no model is perfect), the second stage of verification is to test if this non-conformity significantly affects the model's ability to simulate relevant observed rainfall statistics. Multiple realisations of rainfall are simulated for the gauged time period and sites, not conditional on the observed historical rainfall, while still being conditional on the historical input variables $\mathbf{X}$. This simulation represents the range of possible rainfall time-series which could have occurred (according to the model) given the historic climate variability. If the model is adequate, the observed rainfall data will appear to be one realisation from the simulated distribution of rainfall (Chandler and Wheater, 2002; Yang *et al.*, 2005). Because the observed rainfall statistics have some uncertainty themselves due to the missing data, this stage of verification is preceded by using the model for infilling the historical record, in our case from 1855 to 2011. While the infilled data are dependent on the model itself, and thus not a perfect test-bed, the infilling uncertainty proves to be low in the case study; moreover, explicitly estimating the uncertainty in the historical rainfall in this way is considered an improvement upon the typical practice of neglecting observation uncertainty. The following specific comparisons of simulated and infilled rainfall were used.

(*a*) Time-series of annual site-averaged rainfall, winter (October to March) site-average rainfall, and summer (May to September) site-averaged rainfall. These averages do not include any weighting to represent the area represented by each site.
(*b*) Statistics of inter-annual variability of site-averaged rainfall for each of the 12 months: average, standard deviation, skewness and selected percentiles.
(*c*) Annual average rainfall at each site.
(*d*) Variance, skewness and correlation of annual average rainfall over sites.
(*e*) Two-year, five-year and ten-year running averages of annual and winter rainfall, to assess the ability of the model to represent persistence.

It was decided not to apply split-sample validation, in which some data are omitted from model fitting and used solely for verification, in order to maximise data available for model fitting. However, the analysis of model residuals provides information about model bias over time and space that is similar to split-sample testing, and the testing of the model on various statistics not used in the model fitting is the typical approach to verification of stochastic rainfall models (Chandler and Wheater, 2002; Yang *et al.*, 2005). Although there is not enough space to show them all, a sample of results is shown below. Some supplementary results are available on the article's web-page.

4. Results

4.1 Deterministic predictors of mean rainfall

The input climate variables found to significantly affect the time variation of rainfall in at least some months are: Central England temperature, Mean Sea Level Pressure and the North Atlantic Oscillation index. For most months, a linear trend (increasing rainfall) was also present. This trend term by itself does not necessarily mean increasing rainfall because it is the combined effect of all the input variables that matters; however repeating the regression using only a trend term also illustrates a general increase in rainfall. Easting and northing coordinates and altitude were significant in explaining the regional variability. The coefficient estimates over the 12 months are shown in Figures 3(a) to (i), together with intervals which are not significantly different from zero at the 95% significance level. There is no easy analytical solution for these 95% significance intervals because of

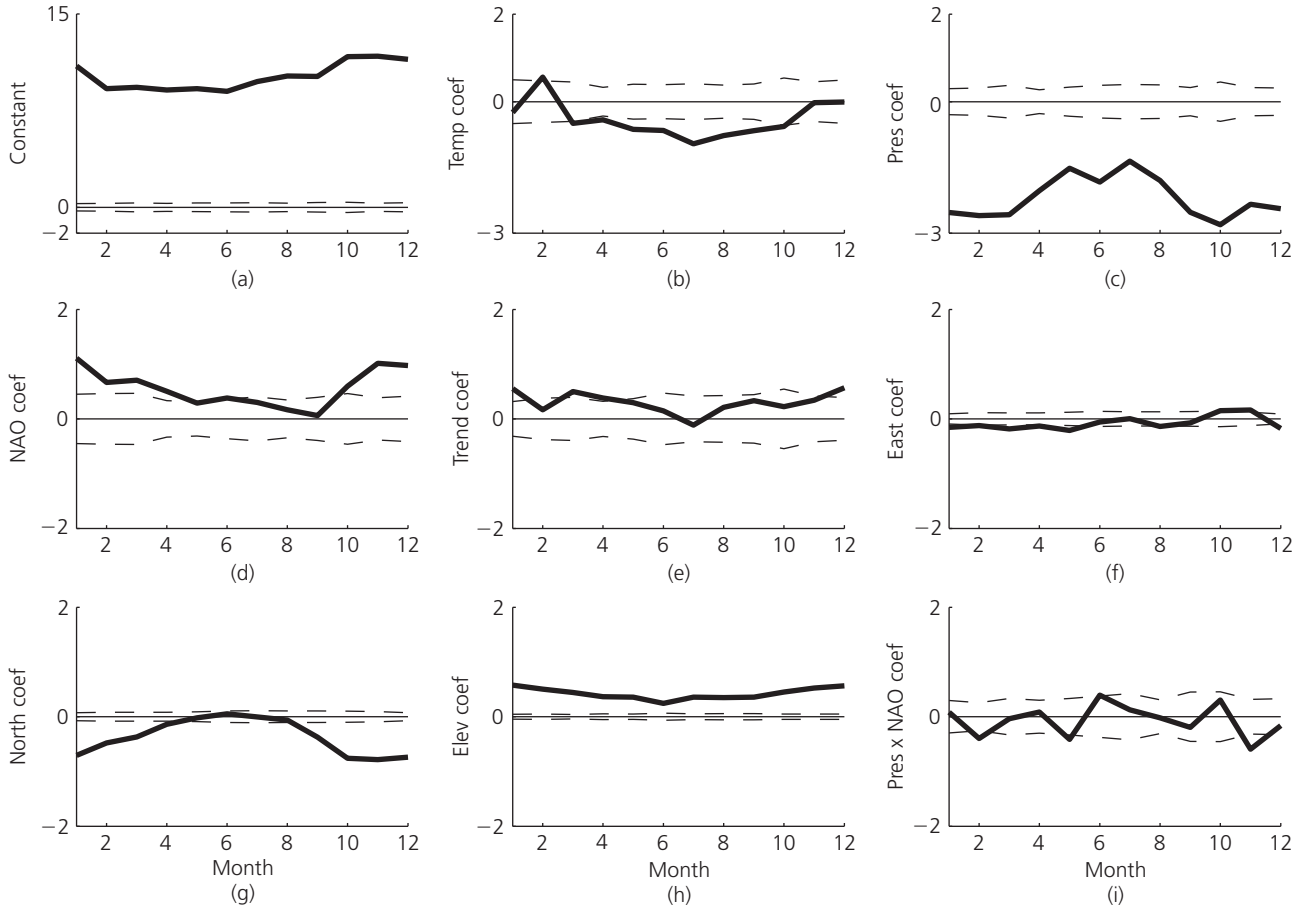

Figure 3. Regression coefficients for each month. Dashed lines represent the interval which is not different from zero at the 95% significance level

the inter-site dependencies, and so they have been estimated using simulation (i.e. with the regression coefficients set to zero, the data were simulated 200 times from the estimated error model, and 200 sample regression coefficients were identified: the top and bottom 2·5% were removed to give the simulated intervals). There was interaction between the effects of coefficients due to the co-linearity between input variables. This was most notable for the coefficients for Mean Sea Level Pressure and Central England temperature (e.g. in January, their correlation was −0·77), and for the coefficients for Central England temperature and trend (e.g. in January, their correlation was −0·25). This leads to relatively high variance in these coefficient estimates and hence wide significance intervals in Figure 3. Nevertheless, Figure 3 illustrates that all the input variables have significant independent effects in at least some months. The second-order effect of variables (e.g. whether the North Atlantic Oscillation has greater influence for the more southerly gauges) was tested by using combinations of variables as inputs to the regression. The only significant second-order effect was the combined effect of Mean Sea Level Pressure and the North Atlantic Oscillation: in February, May and November, pressure had a greater influence when the oscillation was strong (Figure 3(i)). The magnitude of the coefficient values are measures of relative sensitivity of the rainfall to the inputs showing the dominant roles of Mean Sea Level Pressure, northing and altitude (Figures 3(c), (g) and (h)).

The linear trend term is significant at the 95% level in only three months – January, March and December. However, it is above zero for all months except July, and for this to occur due to random variability is extremely improbable. Hence it was concluded that the trend over the period 1855–2011 was significant in all seasons except summer. For the purpose of explaining the rainfall variability and providing the potential for extrapolating the model, the trend would ideally be explained by physical phenomena. Various attempts were made to introduce explanatory variables to explain this trend, including non-linear transforms of Central England temperature, Mean Sea Level Pressure and the North Atlantic Oscillation index and their interactions, but these were not helpful. If time-series of atmospheric carbon dioxide concentrations (constructed from the Hawaii measurements of Keeling *et al.* (1995) and the Antarctic ice-cores of Etheridge *et*

al. (1996)) are used as inputs then the R^2 values are slightly increased and the linear trend term becomes much less significant. While the statistical explanation for this is simple – the carbon dioxide data increase over time hence replacing the trend term – there is no clear physical explanation of why carbon dioxide should explain rainfall variability when the climate variables do not and hence the carbon dioxide input was not adopted. The attraction of this model, however, is noted again below when considering its effect on the structure of errors.

4.2 Error analysis

The regression model summarised in Figure 3 used the same Box–Cox transform for all 12 months and all 50 sites with optimised λ of 0·41. The use of a constant λ for all 12 regression models was necessary to make meaningful comparisons of coefficients between months (the Box–Cox transform rescales the data, so that use of 12 different coefficients would result in coefficients which were not comparable over months as they are in Figure 3). The use of constant λ, however, causes undesirable skewness in the errors for several months (in July, for example, the skewness coefficient was 0·39) and reduces applicability of the error model specified in Equations 3, 4 and 5. For further analysis, therefore, λ was optimised for each month individually, which produced near-normal distributions of errors for each of the 12 models. Optimising λ individually for all 50 sites for each month is possible, but likely to lead to non-unique solutions and, in any case, using a spatially uniform value produced satisfactory error distributions.

When averaged over sites, the errors showed little apparent structure. This included no discernible relationships between errors and input variables, the error histograms had no visible deviation from a zero-mean normal distribution, and there were no significant autocorrelations of errors from one month to the next. The latter result is illustrated in Figure 4, in which the error autocorrelations for the gauges in Kent are plotted. Although there is significant month-to-month correlation in the actual rainfall time series, this is represented by the deterministic part of the model, leaving the month-to-month dependency between errors insignificant. This supports the view that a continuous time series can be simulated using an independent model for each month. There was a tendency for the model to underestimate rainfall in the early years of the record, introducing a visible bias in the errors in the period 1855–1875 (although not illustrated

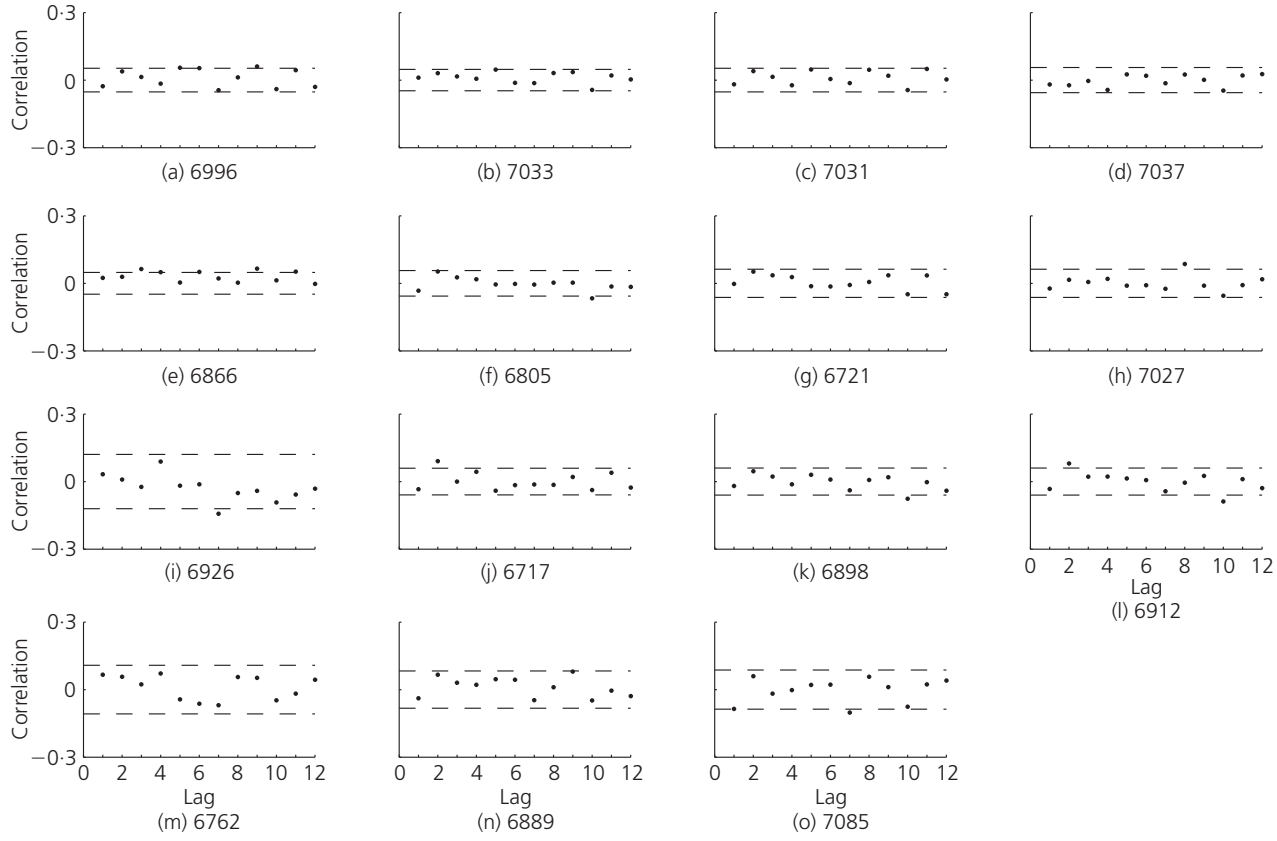

Figure 4. Autocorrelation of errors for the gauges in Kent at lags 0–12 months. Dashed lines represent the interval which is not different from zero at the 95% significance level

here, this will be seen in the verification results described later and shown in Figures 8 and 9). This apparent bias occurred because the linear trend term describing the general increase in winter rainfall was applied over the whole series, whereas closer inspection reveals that there was a much weaker trend between 1855 and 1900. Again, it is tempting to use the atmospheric carbon dioxide concentrations instead of the linear trend: this substantially reduces the bias because carbon dioxide concentrations rose more slowly in the pre-1900 period. However, as previously noted, there is a reluctance to do so without a physical explanation. Also, the small number of gauges operational during these problematic early years (Figure 2) means that there would be relatively low confidence in such a model.

The spatial error analysis also illustrated potential minor flaws in the model. This is seen in Figure 5, which, as an example, plots the mean monthly errors for the 15 sites in Kent. While the *statistical* significance of many of these errors is indicated by their lying outside the estimated 95% significance intervals, their *physical* significance is questionable. This is because the bias may be explained by measurement error, for example Rodda and Smith (1986) present 5% as the typical under-catch associated with gauges not installed at ground level, and they found that in some cases the measurement error was much larger than that. From our model, the maximum observed relative error, out of all sites, was 5% (at the driest site in the region, Figure 5(m)). Nevertheless, an improved spatial model should be considered in future model development.

For each month, there was no evident spatial structure in the error variance estimates. This is illustrated in Figure 6 which shows the sample standard deviation of errors for the sites in Kent with their 95% confidence intervals (the intervals are calculated using the approximate solution described by Kottegoda and Rosso, 2008; p. 244). For comparison, superimposed upon those results as a horizontal line, Figure 6 also shows the sample standard deviation of errors when all 50 sites are considered together, illustrating that, with very few exceptions, this regionally lumped value is a fair estimate for each individual site. Hence the assumption is made that variance of errors is uniform over the whole region within any one month. The correlation of errors between sites, on the other hand, displays a strong spatial structure, with correlation decreasing with distance as described by Equation 6. The fitted correlogram models are illustrated in Figure 7. The models are relatively consistent over the months, with a faster decline in correlation with distance from April to September in comparison with October to March, reflecting the increasing role of more localised, convective events in summer.

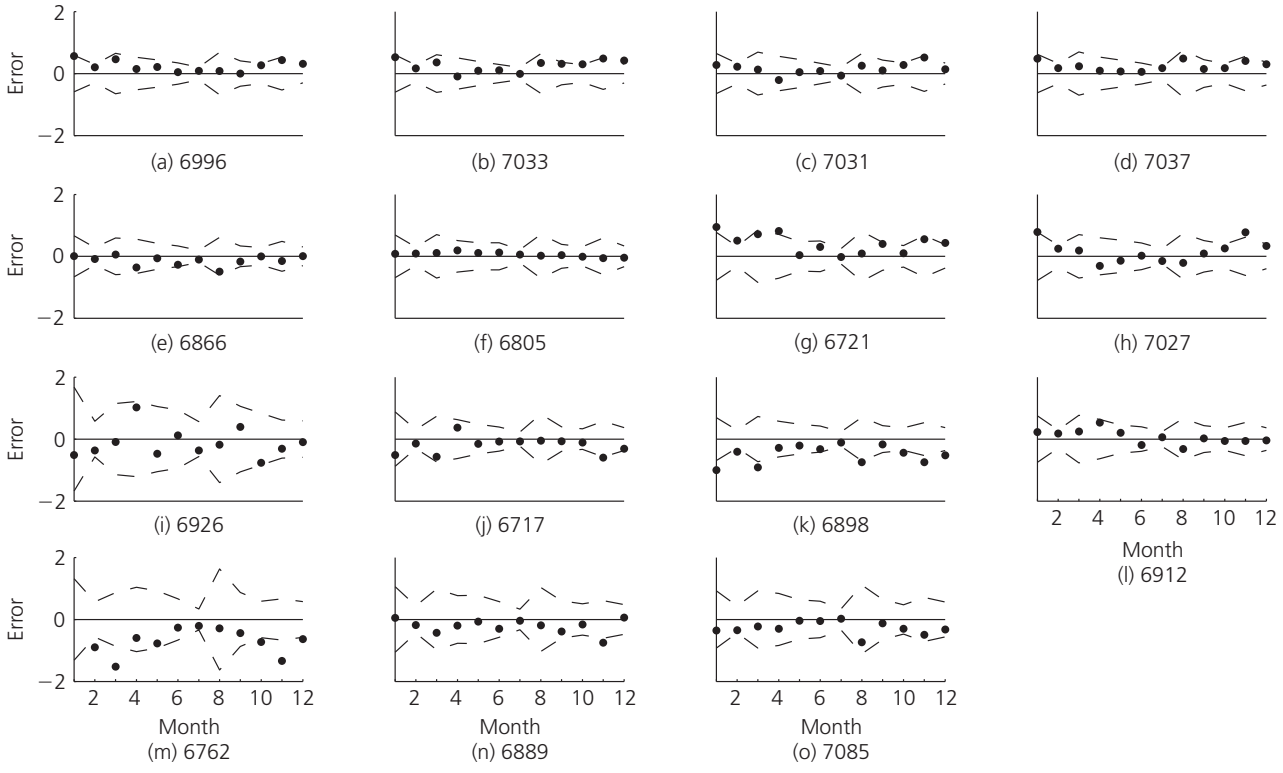

Figure 5. Illustration of bias: errors (in transformed rainfall) averaged over all years for the gauges in Kent. Dashed lines are estimated 95% tolerance intervals around zero

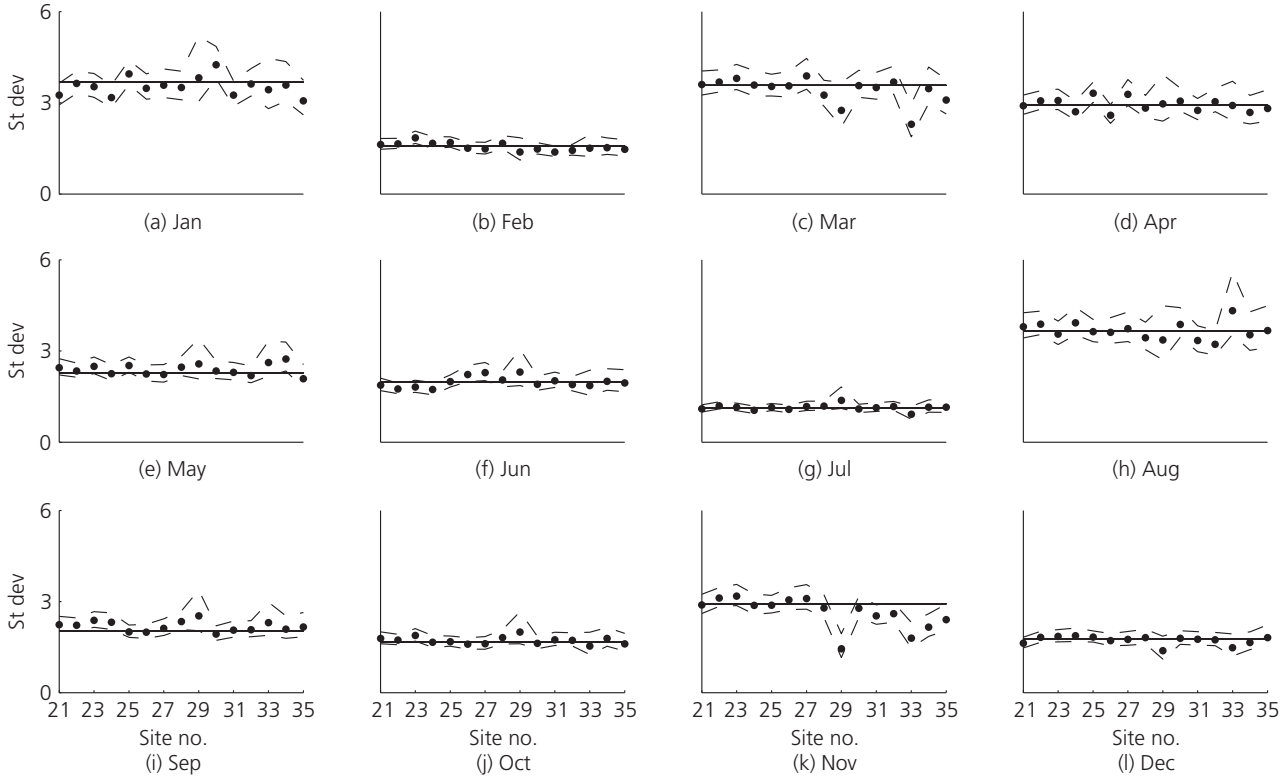

Figure 6. Standard deviation of errors for each site in Kent. Dashed lines are approximate 95% confidence intervals. Horizontal line is the estimate assuming that variance is uniform over all 50 sites in the south-east region. Note: months cannot be inter-compared in this plot because a different Box–Cox transform was used for each month

4.3 Model verification

First, the historical data from 1855–2011 were infilled using the model. For each month/site with missing data, 200 samples of the time-series of errors were used to represent the stochastic variability. The 200 time-series of infilled annual, summer and winter site-average rainfall are shown in Figure 8. Notably, uncertainty in the infilled data is highest during the earlier years when there were fewer active gauges. Nevertheless, the uncertainty is not overriding in terms of the regional rainfall estimate, because: (1) much of the rainfall variability is predictable by the regression equation; (2) the relatively high inter-site correlations evident in Figure 7 mean that the long-term sites provide much of the necessary information about residual variability; and (3) averaging over sites, and over years or seasons (as in this plot) reduces the variance. If considering sub-regions, the uncertainty in the earlier years would become higher especially when moving further from the long-term gauges; and if considering rainfall in individual months then the uncertainty is also higher. The infilling was also applied to synthesised sites on a 5 km grid, producing a spatially quasi-continuous data set covering the period 1855–2011 (results not shown here).

Second, 200 time-series of rainfall (not conditioned upon the observations) were simulated with the model to represent statistically plausible ranges of rainfall. The 95% confidence intervals derived from the ensemble of site-average rainfall are shown in Figure 8, as well as the maximum and minimum values from the ensemble. Comparing the infilled and simulated distributions in Figure 8, it appears that the infilled data are a sample from the simulated rainfall, supporting the view that the model usefully represents the historic variability. The rainfall during the extreme winter drought of 1975–1976 and the extreme summer drought of 1921 are only just encompassed by the simulation bounds, implying that these drought events were extreme given the large-scale climatic conditions at the time. The long drought of 1887–1910 also appears from Figure 8 to be captured by the simulations, as are the dry winters in 1879–1880 and 1897–1898, and the pairs of dry winters in 1995–1997 and 2004–2006.

Figure 8, however, does not allow inter-annual drought persistency to be properly evaluated. To do so, two-year, five-year and ten-year running averages are presented in Figure 9. This illustrates that the series of droughts from 1887 to 1910 are

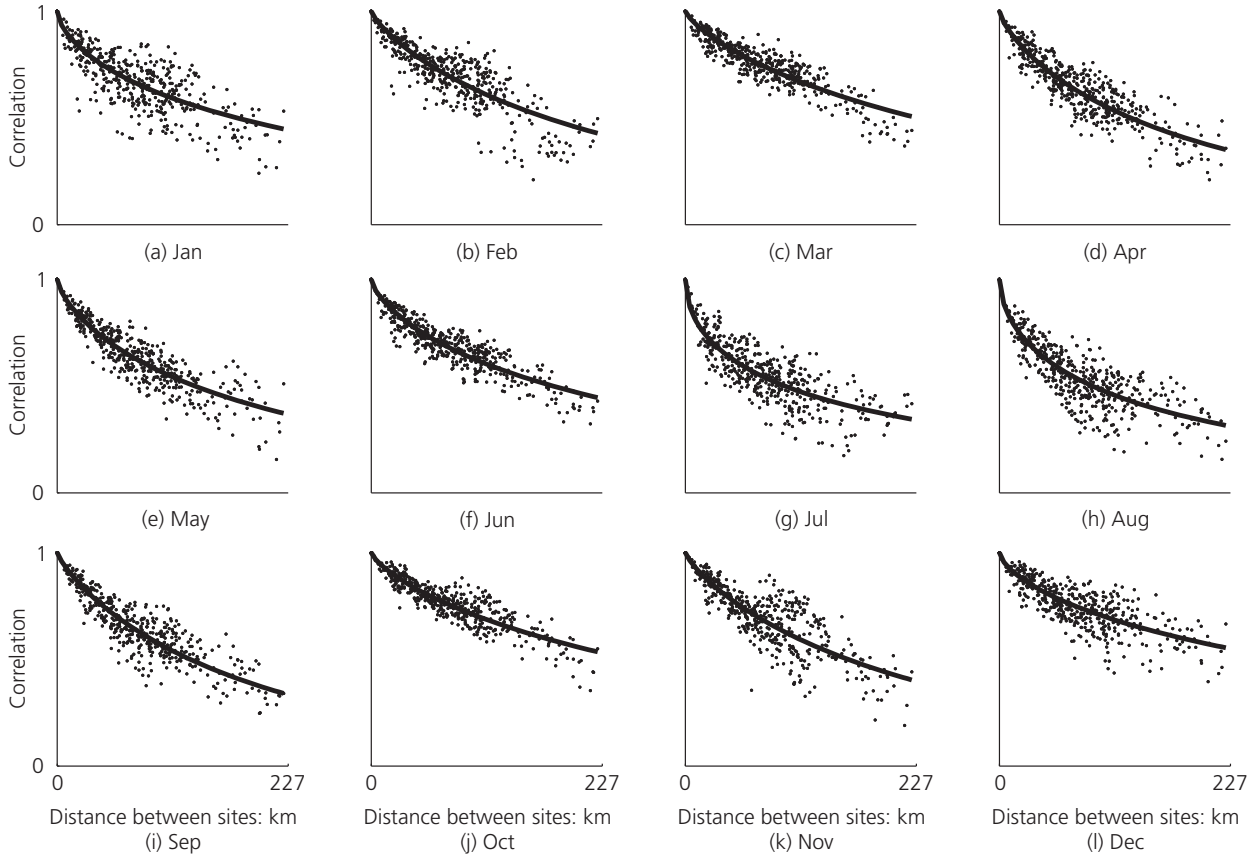

Figure 7. Correlogram describing error spatial correlation as a function of distance between sites

captured, but only by the driest of the simulations, illustrating the extremeness of this drought period given the forcing climate. It is pertinent to note that, according to Figure 9, the most severe two-year drought on record (1920–1922) could recur; indeed it appears that the south-east region was fortunate in 2004–2006 not to have suffered a similar episode given the general climatic conditions at that time. As previously discussed, a feature of Figure 9 is the model's tendency to underestimate multi-year rainfall in the period 1855–1875.

Figure 10 shows a number of temporal and spatial statistics of the infilled and simulated data. Generally, this further supports the view that the model is approximating the properties of the observed rainfall. Some statistics – the minimum, maximum, standard deviation and skewness over time – are persistently towards the lower bound of the simulated distribution, which is expected due to the skewed nature of the rainfall distribution. Figure 10(c) shows, however, a clear tendency to over-estimate the maximum July and October rainfalls: this is associated with errors in representing the distribution of transformed rainfall in these months using a normal distribution. Another interesting result in Figure 10 is the model's tendency to overestimate the spatial skewness of average monthly rainfall in October, November and December. While the model predicts insignificant spatial skewness in these months, the infilled data imply that there are a few sites with much lower monthly averages than the norm, producing significant negative skewness. This is due to the overestimation of rainfall at some of the driest sites in the region, in northern Kent. This was seen in the negative residuals at sites 6762 and 6898 in Figure 5. As previously discussed, this may be resolved by using a more sophisticated spatial model (e.g. quadratic terms for east and north coordinates), however arguably this would be over-fitting as the biases at these sites are within possible measurement errors.

5. Discussion

This paper was motivated by the need for spatially and temporally complete, long-term rainfall records to support regional drought management. The south-east UK is an example of a region which is vulnerable to extreme droughts, and repetition of historical inter-annual droughts is a worrying prospect under current and future demand for water. The south-east UK is, however, fortunate in having gauged sites going back to the mid-nineteenth century, allowing more insight into rainfall variability and more

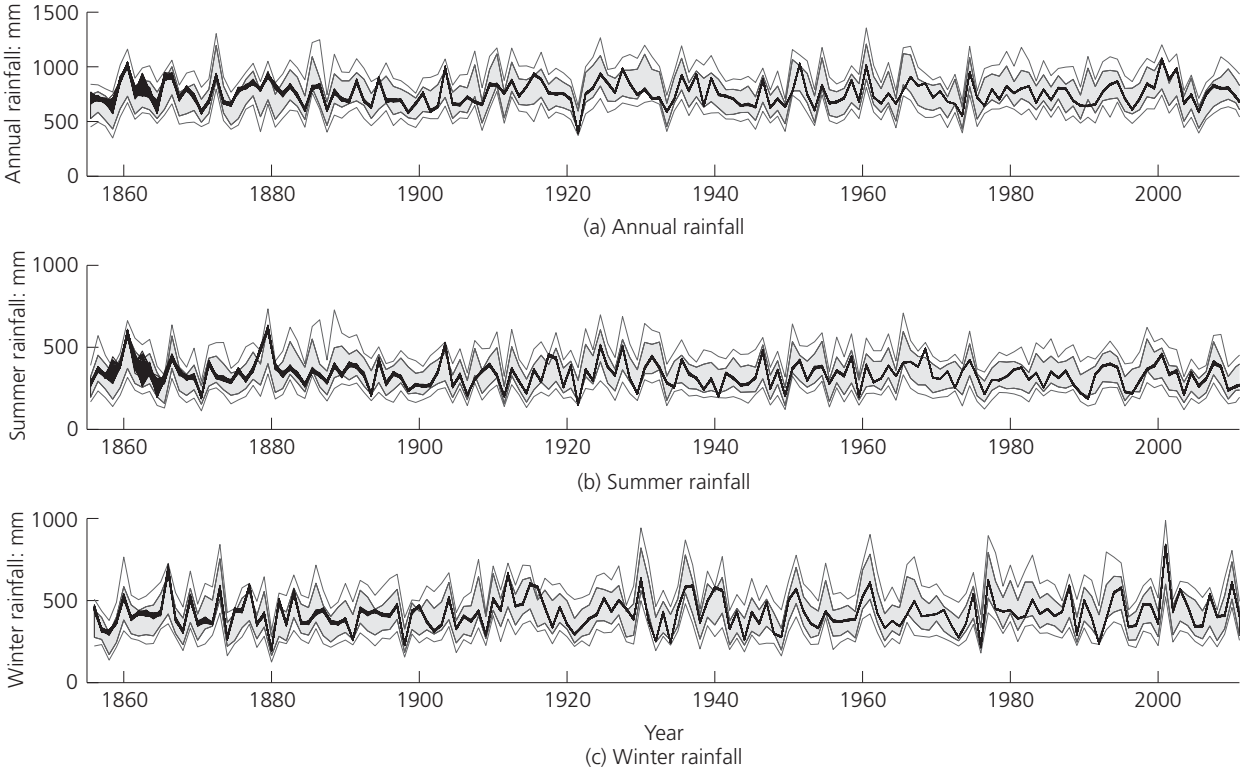

Figure 8. Infilled and simulated site-average rainfall: annual, winter (October to March) and summer (April to September) periods; 95% intervals of ensemble are shaded; outer lines are bounds of ensemble; black lines are ensembles of infilled data

reliable estimation of rainfall extremes than is generally possible. Nevertheless, there are long periods of missing records, and many parts of the region with no long-term records. Significant effort has previously been made under the UKCP09 programme to produce a nationally applicable statistical downscaling tool for UK daily rainfall simulation. However, that downscaling tool is not applicable to infilling historic rainfall in a spatially consistent manner, and may have limitations in replicating extreme historic droughts because it is not linked to the physical drivers of rainfall and has been fitted using a limited range of droughts (Chun, 2011; Jones *et al.*, 2009). The model presented in this paper aims to address these limitations and hence provide complementary data sets.

This paper described a set of regression models for characterising rainfall variability, and infilling and simulating monthly rainfall. The models include a deterministic component that models expected monthly rainfall under specified large-scale climatic conditions, and also a stochastic component that simulates the random variability around the expected value. Gridded rainfall can be produced for a range of observed or synthetic droughts. Using the case study of south-east UK, 50 long-term rain gauges with records spanning from 1855–2011 were used to identify and assess the models. The large-scale variables found to affect rainfall were generally consistent with the findings of previous research on UK rainfall: air pressure, air temperature and North Atlantic Oscillation. A positive linear trend term was identified throughout the twentieth century in all seasons except summer. However, the trend was weak in comparison with the other effects and the random component, and did not preclude recurrence of the severe inter-annual droughts observed in the record.

The model assessment illustrates the potential value of relatively simple rainfall models for generating realistic monthly rainfall patterns. Performance in terms of error diagnosis and comparison of infilled and simulated statistics was considered to be good, although there were two main issues which might benefit from further investigation. First, spatial biases arose from the use of a simple spatial model, causing apparent over-estimation of rainfall at some of the driest sites in Kent. These biases might be explained by rainfall measurement errors, although their particular prevalence in north Kent makes this seems unlikely. Second, temporal biases arose in the period 1855–1875 because the linear trend was weaker in this early period. Using atmospheric carbon dioxide as an input helped to explain the non-stationarity in the trend. It may be speculated that carbon dioxide has influenced

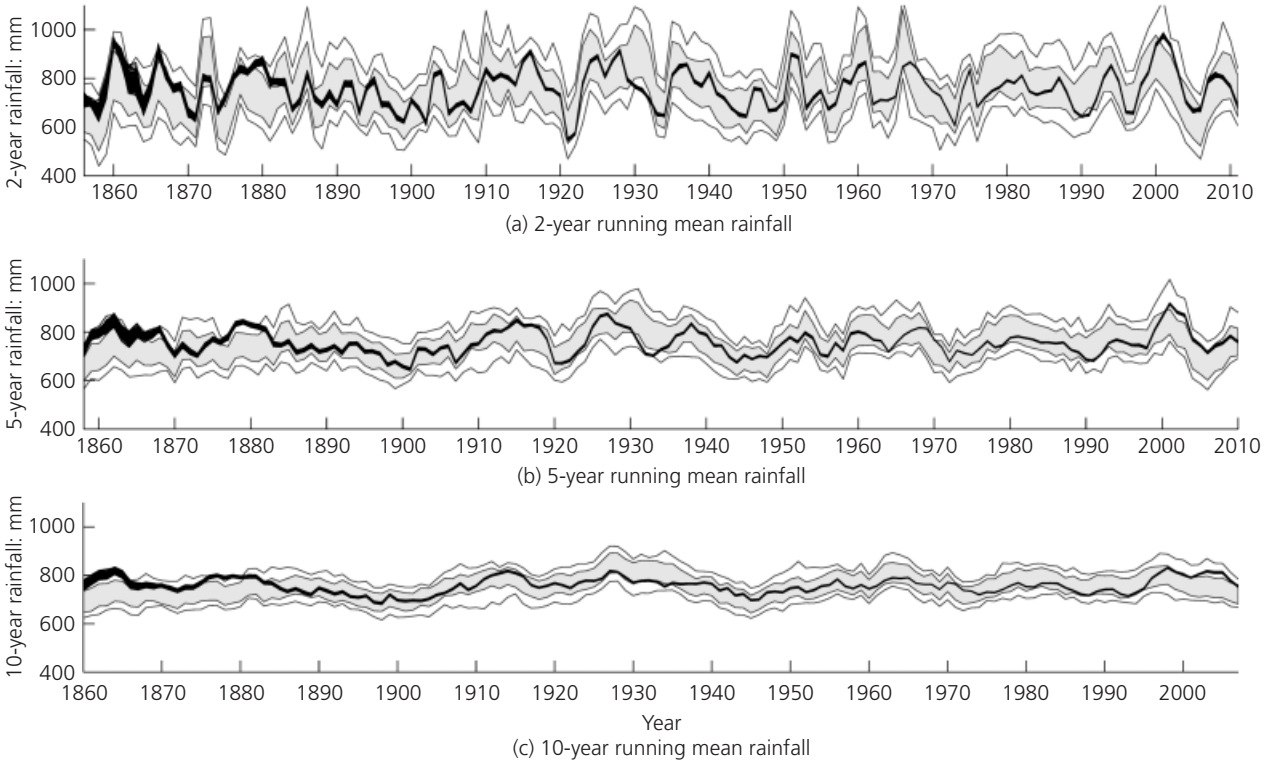

Figure 9. Infilled and simulated 2-year, 5-year and 10-year running means; 95% intervals of ensemble are shaded; outer lines are bounds of ensemble; black lines are ensembles of infilled data

global climate patterns, and hence south-east UK rainfall, in a manner that cannot be represented by the combinations of pressure, temperature and NAO and their interactions investigated in this paper. For example, although east Atlantic 'blocking' patterns are known to be influenced by global climate and to affect rainfall (Pelly and Hoskins, 2003), they were omitted in this investigation because reconstructions of blocking only date back to 1958. This deserves some further investigation. In terms of the model's ability to simulate inter-annual drought, indices of the long droughts within 1887–1910 were within the range of simulations, as were indices of the extreme two-year droughts of 1920–1922, 1933–1934 and 1975–1976. According to the model, the recent droughts of 2004–2006 could have been much more severe given the climatic conditions at the time – potentially more severe than the 1920–1922 event.

The ability of the model to simulate rainfall as a function of large-scale climate variables and indices makes it tempting to employ the model for downscaling global climate model outputs for climate change impacts assessment. However, extrapolating the historic signals to future climate in this manner, although common practice (e.g. Chun *et al.*, 2009; Haylock *et al.*, 2006; Maraun *et al.*, 2010), is not recommended unless it can be shown that the signals are expected to be stationary under a changed climate. Further research is required towards characterising non-stationarity and how it might be resolved in the model. Perhaps the primary limitation of the model described here is that for some applications daily rainfall would be preferred. Development to simulate daily rainfall would require the wet–dry day distribution to be modelled independently of the rainfall depth distribution (Mehrotra and Sharma, 2010). This would naturally lead to the more generalised linear modelling techniques used, for example, by Yang *et al.* (2005). However, for regional analysis of inter-annual droughts in systems with large storage capacity such as south-east UK, monthly scale analysis is likely to be sufficient. Another potential extension to the analysis would be extending records even further back in time by including palaeo data as predictors (Henley *et al.*, 2011).

6. Practical relevance and potential applications

A major challenge in water resource planning is the hindcasting of hydrological data to ensure that possible extreme droughts, including inter-annual sequences of droughts, are adequately considered. A second challenge, important when considering options for intra- and inter-regional water transfers, is spatially consistent characterisation of droughts. These challenges are especially relevant in the water-stressed south-east UK. Currently

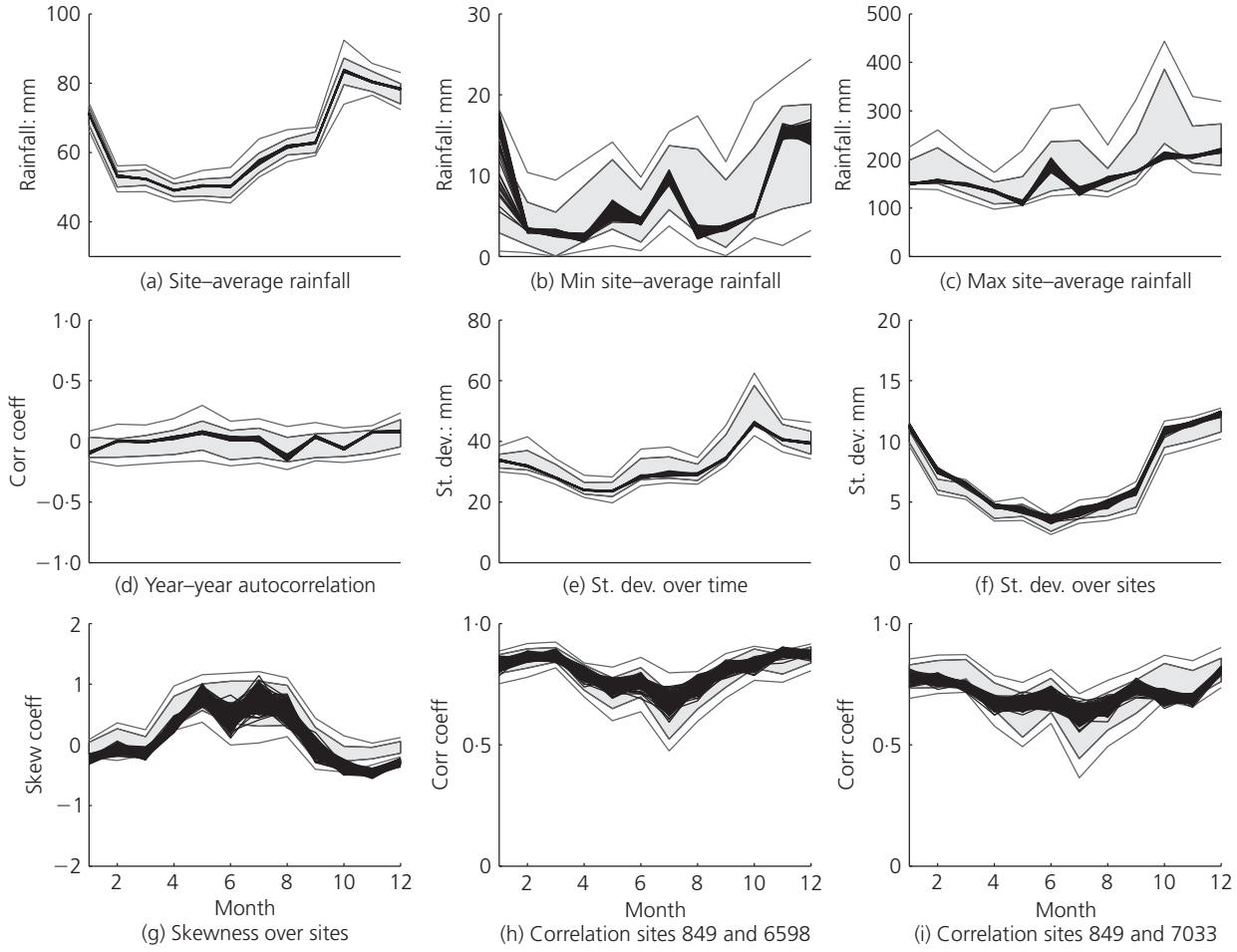

Figure 10. Selected statistics of the infilled and simulated rainfall; 95% intervals of ensemble are shaded; outer lines are bounds of ensemble; black lines are ensembles of infilled data

available climate modelling tools and data sets, such as UKCP09, are not by themselves designed to meet these challenges. This paper describes and tests a statistical model that infills and extends historical rainfall observations to allow improved consideration of extreme and inter-annual droughts in the south-east UK, with potential applicability to other regions where similar problems exist.

Acknowledgements

This research was supported by the Grantham Institute for Climate Change at Imperial College London. Thanks also to NERC and the Meteorological Office, and other data providers listed in Table 1.

REFERENCES

Arnell N and Delaney E (2006) Adapting to climate change: Public water supply in England and Wales. *Climatic Change* **78(2)**: 227–255.

Barnston AG and Livezey RE (1987) Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Monthly Weather Review* **115(6)**: 1083–1126.

Barry RG and Chorley RJ (eds) (2003) *Atmosphere, Weather and Climate*. Routledge, London, UK.

Burke EJ and Brown SJ (2010) Regional drought over the UK and changes in the future. *Journal of Hydrology* **394(3/4)**: 471–485.

Chandler RE and Wheater HS (2002) Analysis of rainfall variability using generalized linear models: A case study from the west of Ireland. *Water Resources Research* **38(10)**: 1192, http://dx.doi.org/10.1029/2001WR000906.

Chun KP (2011) *Statistical Downscaling of Climate Model Outputs for Hydrological Extremes*. PhD dissertation, Imperial College London, UK.

Chun KP, Wheater HS and Onof CJ (2009) Streamflow estimation for six UK catchments under future climate scenarios. *Hydrology Research* **40(2–3)**: 96–112.

Draper NR and Smith H (eds) (1998) *Applied Regression Analysis*. Wiley, New York, USA.

Etheridge DM, Steele LP, Langenfelds RL, Francey RJ, Barnola JM and Morgan VI (1996) Natural and anthropogenic changes in atmospheric CO_2 over the last 1000 years from air in Antarctic ice and firn. *Journal of Geophysical Research* **101(2)**: 4115–4128.

Fowler HJ, Blenkinsop S and Tebaldi C (2007) Linking climate change modelling to impacts studies: Recent advances in downscaling techniques for hydrological modelling. *International Journal of Climatology* **27(12)**: 1547–1578.

Hanssen-Bauer I and Førland EJ (1998) Long-term trends in precipitation and temperature in the Norwegian Arctic: Can they be explained by changes in atmospheric circulation patterns? *Climate Research* **10(2)**: 143–153.

Haylock MR, Cawley GC, Harpham C, Wilby RL and Goodess CM (2006) Downscaling heavy precipitation over the United Kingdom: a comparison of dynamical and statistical methods and their future scenarios. *International Journal of Climatology* **26(10)**: 1397–1415.

Henley BJ, Thyer MA, Kuczera G and Franks SW (2011) Climate-informed stochastic hydrological modeling: Incorporating decadal-scale variability using paleo data. *Water Resources Research* **47(11)**: W11509, http://dx.doi.org/10.1029/2010WR010034.

Horn RA and Johnson CR (1985) *Matrix Analysis*. Cambridge University Press, Cambridge, UK.

Hulme M and Barrow E (eds) (1997) *Climates of the British Isles: Present, Past and Future*. Routledge, London, UK.

Jenkins GJ, Perry MC and Prior MJ (2008) *The Climate of the United Kingdom and Recent Trends*. Met Office Hadley Centre, Exeter, UK.

Jones PD, Kilsby CG, Harpham C, Glenis V and Burton A (2009) *UK Climate Projections Science Report: Projections of Future Daily Climate for the UK from the Weather Generator*. University of Newcastle, Newcastle Upon Tyne, UK.

Keeling CD, Whorf TP, Wahlen M and van der Plicht J (1995) Interannual extremes in the rate of rise of atmospheric carbon dioxide since 1980. *Nature* **375(6533)**: 666–670.

Kenabatho PK, McIntyre N, Chandler RE and Wheater HS (2011) Stochastic simulation of rainfall in the semi-arid Limpopo basin. *International Journal of Climatology* **32(7)**: 1112–1127.

Kigobe M, McIntyre N, Wheater HS and Chandler RE (2011) Multi-site stochastic modelling of rainfall in Uganda. *Hydrological Sciences Journal* **56(1)**: 17–33.

Kottegoda NT and Rosso R (2008) *Applied Statistics for Civil and Environmental Engineers*. Blackwell, UK.

Lavers D, Prudhomme C and Hannah DM (2010) Large-scale climatic influences on precipitation and discharge for a British river basin. *Hydrological Processes* **24(18)**: 2555–2563.

Maraun D, Wetterhall F, Ireson AM *et al.* (2010) Precipitation downscaling under climate change. Recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics* **48(3)**: RG3003, http://dx.doi.org/10.1029/2009RG000314.

Marsh TJ (1996) The 1995 UK drought – a signal of climatic instability. *Proceedings of the Institution of Civil Engineers – Water Maritime and Energy* **118(3)**: 189–195.

Marsh T, Cole G and Wilby R (2007) Major droughts in England and Wales, 1800–2006. *Weather* **62(4)**: 87–93.

McIntyre N, Lees M, Wheater HS, Onof C and Connorton B (2003) An approach to the evaluation and visualisation of risk to security of water resources. *Proceedings of the Institution of Civil Engineers – Water and Maritime Engineering* **156(1)**: 1–11.

Mechler R and Kundzewicz ZW (2010) Assessing adaptation to extreme weather events in Europe – Editorial. *Mitigation and Adaptation Strategies for Global Change* **15(7)**: 611–620.

Mehrotra R and Sharma A (2010) Development and application of a multisite rainfall stochastic downscaling framework for climate change impact assessment. *Water Resources Research* **46(7)**: W07526, http://dx.doi.org/10.1029/2009WR008423.

Murphy SJ and Washington R (2001) United Kingdom and Ireland precipitation variability and the North Atlantic sea-level pressure field. *International Journal of Climatology* **21(8)**: 939–959.

Pelly JL and Hoskins BJ (2003) A new perspective on blocking. *Journal of the Atmospheric Sciences* **60(5)**: 743–755.

Phillips ID and McGregor GR (2002) The relationship between monthly and seasonal south-west England rainfall anomalies and concurrent North Atlantic sea surface temperatures. *International Journal of Climatology* **22(2)**: 197–217.

Rodda JC and Smith SW (1986) The significance of the systematic error in rainfall measurement for assessing wet deposition. *Atmospheric Environment* **20(5)**: 1059–1064.

Searle SR (1971) *Linear Models*. Wiley, New York, USA.

Segond ML, Onof C and Wheater HS (2006) Spatial–temporal disaggregation of daily rainfall from a generalized linear model. *Journal of Hydrology* **331(3–4)**: 674–689.

Smith JT, Clarke RT and Bowes MJ (2010) Are groundwater nitrate concentrations reaching a turning point in some chalk aquifers? *Science of the Total Environment* **408(20)**: 4722–4732.

Subak S (2000) Climate change adaptation in the UK water industry: Managers' perceptions of past variability and future scenarios. *Water Resources Management* **14(2)**: 137–156.

Tabachnik BG and Fidell LS (1996) *Using Multivariate Statistics*. Harper Collins, NewYork, USA.

Thyer M, Kuczera G and Wang QJ (2002) Quantifying parameter uncertainty in stochastic models using the Box–Cox transformation. *Journal of Hydrology* **265(1–4)**: 246–257.

Uppala SM, Kållberg PW, Simmons AJ *et al.* (2005) The ERA-40 Reanalysis. *Quarterly Journal of the Royal Meteorological Society* **131(612)**: 2961–3012.

von Lany PH, Hepworth N, Hawker PJ and Choudhury F (2008) Working towards integrated and more sustainable water resources planning in Southeast England. *Proceedings of the BHS 10th National Hydrology Symposium, Exeter*, pp. 1–8.

Wilby RL, Wedgbrow CS and Fox HR (2004) Seasonal predictability of the summer hydrometeorology of the River Thames, UK. *Journal of Hydrology* **295(1–4)**: 1–16.

Xu CY (1999) From GCMs to river flow: A review of

downscaling methods and hydrologic modelling approaches. *Progress in Physical Geography* **23(2)**: 229–249.

Yang C, Chandler RE, Isham VS and Wheater HS (2005) Spatial-temporal rainfall simulation using generalized linear models. *Water Resources Research* **41(11)**: 11415.

Zaidman MD, Rees HG and Young AR (2002) Spatio-temporal development of streamflow droughts in northwest Europe. *Hydrology and Earth System Sciences* **6(4)**: 733–751.