Southern African Monthly Rainfall Variability: An Analysis Based on Generalized Linear Models

CHIARA AMBROSINO
Department of Geography, University College London, London, United Kingdom

RICHARD E. CHANDLER
Department of Statistical Science, University College London, London, United Kingdom

MARTIN C. TODD
Department of Geography, University of Sussex, Brighton, United Kingdom

(Manuscript received 7 July 2010, in final form 23 November 2010)

ABSTRACT

Southern Africa is characterized by a high degree of rainfall variability, affecting agriculture and hydrology, among other sectors. This paper aims to investigate such variability and to identify stable relationships with its potential drivers in the climate system; such relationships may be used as the basis for the statistical downscaling of climate model outputs, for example. The analysis uses generalized linear models (GLMs). The GLMs are fitted to twentieth-century observational data for the period 1957–2006 to characterize the dependence of monthly precipitation occurrences and amounts upon the climate indicators of interest. In contrast with many of the analyses that have previously been used to investigate controls on precipitation in the region, GLMs allow for the investigation of the relationships between different components of the climate system (geographical and climatic drivers) simultaneously. Six climate factors were found to drive part of the rainfall variability in the region, and their modeled effect upon rainfall occurrences and amounts resulted in general agreement with previous studies. Among the retained indices, relative humidity and El Niño accounted for the highest degree of explained variability. The location and intensity of the jet stream were also found to have a statistically significant and physically meaningful effect upon rainfall variability.

1. Introduction

Southern Africa is particularly prone to adverse effects of changes in the climate system (Christensen et al. 2007). The vulnerability of the subcontinent is exacerbated by its high level of dependency on natural and agricultural resources for water, food, and income, and also by its lower ability to adapt to such changes because of multiple stresses, such as extensive poverty and political instability. In addition, future global warming may cause intensification of the hydrological cycle globally (Huntington 2006), with changes in frequencies and durations of periods of above- and below-normal rainfall. Precipitation in the region is characterized by a strong seasonal cycle with a well-defined wet season (December–April) over most of the subcontinent (Reason et al. 2006). Anomalously drier periods during the wet season compromise the availability of water for rain-fed agriculture, upon which much of the population is heavily dependent. Delays in the onset of the rainy season, dry spells during the wet season, and reduction in the length of the growing season (see Brinkmann 1979, for definitions) are some of the factors that contribute to a deficit in water availability for agriculture, thus the failure of a crop. To reduce the vulnerability, provision of climate information is necessary for planning adaptation and mitigation measures (Allan and Soden 2008).

This study aims to investigate the relationships between southern African rainfall and large-scale climate factors, building upon the current knowledge of the structures controlling subcontinental rainfall variability. The

Corresponding author address: Chiara Ambrosino, Department of Statistical Science, University College London, 1-19 Torrington Place, WC1E 6BT, United Kingdom.
E-mail: c.ambrosino@ucl.ac.uk

DOI: 10.1175/2010JCLI3924.1

© 2011 American Meteorological Society
identified relationships can be used for the downscaling of general circulation model (GCM) outputs, bridging the gap between the coarse resolution of those climate models and the required resolution for impact assessment studies (Hewitson and Crane 1996; Wilby and Wigley 1997; Marau et al. 2010). The analysis is based on generalized linear models (GLMs), which allow all the factors to be assessed simultaneously (Chandler 2005). As far as we are aware, this is the first study to take such an approach; most previous studies have used correlation-based analysis to examine the effects of individual factors in isolation. In the next section, a brief review on the previous knowledge about factors affecting southern African rainfall is presented. Section 3 illustrates the selected area of study. The methodology is presented in section 4, and the results are given and discussed in sections 5 and 6. Section 7 gives the conclusions.

2. Background

Southern African rainfall variability has recently been the subject of much research activity at different time scales, from daily (Todd and Washington 1999) to interannual (Jury et al. 1992; Jury 1997; Rocha and Simmonds 1997a; Hoerling et al. 2005; Giannini et al. 1998) to decadal and millennial (Tyson et al. 2002). See Mason and Jury (1997) for an extensive review. Such variability has been explained in terms of variations in pressure and atmosphere circulation patterns (Tyson 1981). Many studies have emphasized the influence of sea surface temperatures (SSTs) on southern African rainfall variability. In particular, El Niño–Southern Oscillation (ENSO) is considered the dominant interannual mode of variability in the tropics and has a clear marked influence upon interannual rainfall variability over the subcontinent (Nicholson and Kim 1997; Mason 2001; Reason and Jagadheesha 2005b; Giannini et al. 2008). Large-scale warming of the equatorial eastern and central Pacific is frequently associated with drought over much of southern Africa (Mason and Jury 1997; Mason 2001; Richard et al. 2001; Jury et al. 2004; Reason and Jagadheesha 2005a; Giannini et al. 2008; Schubert et al. 2009; Findell and Delworth 2010). In particular, El Niño impacts are thought to be most pronounced in the southeast of the continent (Matarira 1990; Rocha and Simmonds 1997b) from December to April. In addition, positive correlations have been found between southern Africa summer rainfall and SST anomalies around the subcontinent, in both the Benguela Current system along the west coast (correlation range 0.31–0.35) and the Agulhas Current in the Indian Ocean (correlation range 0.35–0.40) (Walker 1990). With warm waters, the Indian Ocean is considered the major source of moisture over the subcontinent. Hoerling et al. (2006) linked warming and enhanced convection over the entire Indian Ocean with subsidence and drought over southern Africa during the austral summer months. With the use of GCM simulations, they showed a rainfall reduction across the subcontinent to a specified 1°C Indian Ocean sea surface warming, during the February–April season, up to 100 mm maximum.

Washington and Preston (2006) identified a dipole pattern in the southwest Indian Ocean (SWIO) (warmer conditions in the southwest and cooler in the northwest) linked with extreme summer rainfall events over the subcontinent (positive rainfall anomalies around 1–4 mm day$^{-1}$). Reason (2001) linked the east–west Indian Ocean dipole (IOD) zonal mode (ZM) Saji et al. (1999) to increased moisture advection over southern Africa when SSTs increase from east to west. Recent investigations have also considered the Atlantic Ocean as a source of southern African variability, both for southwest cape winter rainfall and for summer rainfall anomalies, particularly over Angola and Namibia (Reason and Jagadheesha 2005b; Reason et al. 2006). Williams et al. (2008), investigating the association between southern African rainfall and atmospheric and oceanic circulations, identified a combination of anomalously cold SSTs in the central South Atlantic and warm SSTs off the southwestern coast associated with an increase in rainfall extremes over the subcontinent.

Other large-scale features of the climate system have been associated with variations in the southern African rainfall regime. The Brandon–Marion Index (BMI), indicative of changes in the pressure field over the southwest Indian Ocean, has been found correlating with south-eastern African summer rainfall (Rocha and Simmonds 1997a). In particular, Rocha and Simmonds (1997a) found the highest correlation in an area northeast of South Africa during the summer months (December–January).

The quasi-biennial oscillation (QBO) is thought to modulate the effect of ENSO over southern African rainfall (Jury et al. 1994). Reason and Rouault (2005) found a relationship between the positive phase of the southern annular mode (SAM) (Rogers and van Loon 1982) and drier winter conditions over southwestern South Africa. Correlating the time series of the SAM index and a South African rainfall index, they found values as high as 0.4. However, the positive phase of the SAM has been associated with anomalously wet conditions over much of the rest of South Africa, with an increase in precipitation around 0.5 mm day$^{-1}$ (Gillett et al. 2006). The Madden–Julian oscillation (MJO), a mode of intra-seasonal time-scale variability, is characterized by an eastward propagation of tropical deep convection clusters from the Indian Ocean to the western Pacific.
The MJO has been found to significantly influence rainfall in the east and south of southern Africa (Pohl et al. 2007). The North Atlantic Oscillation (NAO) has been linked by McHugh and Rogers (2001) to precipitation variability along and north of the southeastern Africa convergence zone. In the study they described anomalously high (low) convective rainfall occurring over southeast Africa when the NAO is weak (strong). Besides variability at intraseasonal and interannual time scales, Reason and Rouault (2002) linked the Pacific decadal oscillation (PDO) to South African rainfall. They noticed that the warm (cool) phase of the interdecadal variability in the Pacific and Indian Oceans is associated with decreased (increased) rainfall over South Africa (correlation coefficients range −0.4 to −0.53). Ward (1998) linked the near-global-scale Atlantic multidecadal oscillation (AMO) to the multidecadal dry condition over sub-Saharan Africa. Finally, the displacement and change in strength of the jet stream, a belt of strong upper-level winds lying above regions of subtropical high pressure, has been related to rainfall variability over the subcontinent by Richard et al. (2001). According to the study, a strengthening and a northward displacement of the jet stream could lead to a displacement of the cyclonic systems tracks farther north, moistening the southern latitude of the continent, decreasing the easterly wind component, and reducing the tropical temperate troughs over the central southern Africa.

The above-mentioned studies used various station data, gridded products, or climate model outputs. The techniques used to identify the relationships between precipitation and potential drivers comprise correlations between index time series and precipitation as well as simple linear regressions and more qualitative analysis. Composite analyses were used to investigate the conditional probability between data samples recorded during a specific event when a certain other event has or will also take place. In some papers atmospheric GCM simulations were run with idealized sea surface temperature forcing to determine the rainfall response. Other techniques include principal component analysis, canonical correlation analysis, and wavelet decomposition. These are only some of the main techniques used so far. While some of them may be useful tools to identify relationships among atmospheric and oceanographic field patterns, they all focus on individual relationships. Interpretation of these techniques can be problematic when (as seems likely in practice) some of the potential drivers of interest are themselves correlated: in this case, it is likely that several potential drivers may appear to show relationships with precipitation but that only a few of these relationships are causal. To investigate this, it is necessary to examine the effect of all the factors simultaneously; this paper attempts this for the first time.

3. Area considered and data used

In the current study, southern Africa has been defined as the area lying between 12° and 40°S and 0° and 42°E. Twentieth-century precipitation data are derived from the Climatic Research Unit (CRU) time series (TS) 3.0. The dataset is an observed gridded land surface precipitation product at monthly temporal resolution and 2.5° × ~2.5° spatial resolution covering the global land surface. It has been created by interpolation from a nonregular network of stations from different sources. See Mitchell and Jones (2005) for a description of the station sources used to construct the database and for the procedure in deriving the gridded product. Gridded datasets are widely used in climate research because they provide values of the variable of interest at any spatial location and a result especially useful for those areas characterized by a low and sparse density of meteorological stations, such as the African continent (Washington et al. 2006). Moreover, they are available for a relative long period of time, only matched by gauge data. In particular, a number of previously published papers used the same kind of dataset to study the African climate (Jury et al. 2004; Conway et al. 2009; Shongwe et al. 2009). Of course, the gridding procedure may introduce artifacts due to the locations of available stations and the chosen interpolation method. However, with one exception our view is that these gridded data are suitable for characterizing the large-scale regional structures and relationships that are the focus of the present paper. The exception relates to the grid cells centered over Angola, for which no station data were available after the year 1972. Therefore, data from these grid cells, from 1973 onward, have been excluded from the modeling process.

Data on a large number of potential drivers for southern African precipitation have also been used. Table 1 summarizes all the climate variables used in the study, data sources, and data references. All the time series are monthly. These time series reflect the dominant indices identified in literature.

The relative humidity data (RHUM) used here are reanalysis data from the National Oceanic and Atmospheric Administration (NOAA)’s National Centers for Environmental Prediction (NCEP). Although homogeneous techniques are used to generate the reanalysis data, the effect of changes in the atmospheric observing system is unavoidable. One prominent impact has been the introduction, from 1979 to present, of satellite data, particularly over the Southern Hemisphere, causing inhomogeneities in the form of step changes between pre- and post-1979 reanalysis data (Kistler et al. 2001; Sterl 2004; Tennant 2004). To remove the effects both of these inhomogeneities and of seasonality, we used generalized
additive models (Bowman and Azzalini 1997). At each grid cell, a model of the form
\[
\text{RHUM}_t = a + s(t) + \left[ b \times I(\text{Year} < 1979) \right] + \left[ c \times \cos \left( \frac{2\pi \times \text{Month}}{12} \right) \right] + \left[ d \times \sin \left( \frac{2\pi \times \text{Month}}{12} \right) \right] + \epsilon_t
\]
was fitted to the original time series. Here, the trend term \( s(t) \) is a smooth function of time, represented nonparametrically so as to avoid imposing artificial structures that may be unrealistic; the cosine and sine terms represent the seasonal cycle; and \( I(\text{Year} < 1979) \) is an indicator variable taking the value 1 for all observations prior to 1979 and 0 thereafter. Estimation of the smooth trend \( s(t) \), along with the coefficients \( a, b, c, \) and \( d \), was carried out using the `gam()` routine in R.

TABLE 1. Climate drivers of southern African rainfall variability considered in the modeling process.

<table>
<thead>
<tr>
<th>Index</th>
<th>Data reference</th>
<th>Data source</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>El Niño index (Niño)</td>
<td>Rayner et al. (2003)</td>
<td>Met Office Hadley Centre</td>
<td>Normalized SSTs for Niño-3.4 region</td>
</tr>
<tr>
<td>Southern Oscillation index (SOI)</td>
<td>Ropelewski and Jones (1987)</td>
<td>CRU of the University of East Anglia (UEA)</td>
<td>Normalized pressure difference between Tahiti and Darwin</td>
</tr>
<tr>
<td>QBO</td>
<td>Calculated at the Physical Sciences Division (PSD) from NCEP–National Center for Atmospheric Research (NCAR) reanalysis</td>
<td>NOAA/Earth System Research Laboratory (ESRL) PSD</td>
<td>Zonal average of the 30-mb zonal wind at the equator</td>
</tr>
<tr>
<td>NAO</td>
<td>Jones et al. (1997)</td>
<td>CRU-UEA</td>
<td>Normalized pressure difference between a station on the Azores and one on Iceland</td>
</tr>
<tr>
<td>IOD</td>
<td>Saji et al. (1999)</td>
<td>Met Office Hadley Centre</td>
<td>Anomalous SST gradient between the western equatorial Indian Ocean and the southeastern equatorial Indian Ocean</td>
</tr>
<tr>
<td>Indian Ocean sea surface temperatures</td>
<td>Smith et al. (2008)</td>
<td>NOAA/National Climatic Data Center</td>
<td>Normalized SST over the Indian Ocean ((-20^\circ-20^\circ N, 40^\circ-110^\circ E))</td>
</tr>
<tr>
<td>SHT</td>
<td>Brohan et al. (2006)</td>
<td>Met Office Hadley Centre</td>
<td>Standardized Southern Hemisphere surface temperature</td>
</tr>
<tr>
<td>SAM</td>
<td>Marshall (2003)</td>
<td>Station data</td>
<td>MSLP difference between 40° and 65°S</td>
</tr>
<tr>
<td>SWIO</td>
<td>Smith et al. (2008)</td>
<td>NOAA/National Climatic Data Center</td>
<td>Extended reconstruction sea surface temperature differences between northwest and southwest Indian Ocean</td>
</tr>
<tr>
<td>BMI</td>
<td>Kalnay et al. (1996)</td>
<td>NOAA/ESRL PSD</td>
<td>Principal component of south Indian Ocean pressure representing differences between St. Brandon and Marion Islands</td>
</tr>
<tr>
<td>Agulhas sea surface temperatures</td>
<td>Smith et al. (2008)</td>
<td>NOAA/National Climatic Data Center</td>
<td>Leading principal component for sea surface temperature in the Agulhas region</td>
</tr>
<tr>
<td>Benguela sea surface temperatures</td>
<td>Smith et al. (2008)</td>
<td>NOAA/National Climatic Data Center</td>
<td>Leading principal component for sea surface temperature in the Benguela region</td>
</tr>
<tr>
<td>South Atlantic SSTs</td>
<td>Smith et al. (2008)</td>
<td>NOAA/National Climatic Data Center</td>
<td>Leading principal component for sea surface temperature in the South Atlantic Ocean</td>
</tr>
<tr>
<td>Southern African relative humidity (RHUM)</td>
<td>Kalnay et al. (1996)</td>
<td>NOAA/ESRL PSD</td>
<td>Leading principal component of relative humidity over southern Africa ((-35^\circ-0^\circ N, 12.5^\circ-42^\circ E))</td>
</tr>
<tr>
<td>PDO</td>
<td>Zhang et al. (1997)</td>
<td>Met Office Hadley Centre</td>
<td>Leading principal component of North Pacific monthly sea surface temperature variability</td>
</tr>
<tr>
<td>AMO</td>
<td>van Oldenborgh et al. (2009)</td>
<td>Met Office Hadley Centre</td>
<td>Sea surface temperatures in the northern Atlantic</td>
</tr>
<tr>
<td>Jet stream intensity and position</td>
<td>Kalnay et al. (1996)</td>
<td>NOAA/ESRL PSD</td>
<td>Intensity and position of jet stream at 200 mb</td>
</tr>
</tbody>
</table>
The coefficient $b$ is the magnitude of the step associated with the introduction of satellite data. A deseasonalized and desteped series can thus be defined as

$$RHUM_t^a = a + s(t) + [b \times I(Year < 1979)]$$

this is referred to as "adjusted relative humidity." The process is illustrated for a single grid cell in Fig. 1.

As an index of the Southern Hemisphere temperature (SHT), standardized temperature anomalies with respect to the 1961–1990 climatology have been used. The effect of the jet stream upon precipitation in a particular location can be expected to depend on the jet stream intensity and on its position relative to the location of interest. We used 200-mb wind speed reanalysis data to define the intensity and location of the jet stream, corrected for inhomogeneities in the same way as for the relative humidity field. At a particular time $t$, we defined the jet stream position (POS$_t$) to be the latitude corresponding to the maximum zonally averaged winds over the region ($60^\circ$, $20^\circ$S) $\times$ ($0^\circ$, $45^\circ$E). The corresponding zonally averaged maxima intensity were then extracted to form a time series, which is then standardized to form an index (say, INT$_t$) of jet stream intensity.

Because of data availability, the study focused on the second half of the twentieth century. Specifically, 42 yr, from 1957 to 2006, of contemporary records were used in the study.

4. Methodology

a. Generalized linear model for monthly rainfall

To explore the simultaneous effect of the many factors that have been suggested as potential controls upon precipitation in southern Africa, the analysis in this paper is based on GLMs, which can be regarded as an extension of multiple regression techniques. GLMs have been used widely in statistical science and were recently explored in some papers in climate research (e.g., Chandler 2005; Furrer and Katz 2007; Fealy and Sweeney 2007). Chandler and Wheater (2002) and Yan et al. (2002) demonstrated the power of GLMs in analyzing and understanding the relationships among components of the climate system and in quantifying changes in the system. The flexibility of GLMs, coupled with the fact that they are easily interpretable and computationally inexpensive, supports their application in the present study. Here, a GLM for monthly gridded precipitation data is defined by setting up a probability distribution for each time point and grid cell from which the corresponding observation is considered to be drawn. These individual distributions are themselves constructed from the values of other variables (referred to as covariates) specific to that time point and grid cell.

In the GLM framework, the observations are regarded as realized values of a vector of random variables, say, $Y = (Y_1, \ldots, Y_n)'$, all assumed to be generated from the same family of distributions (e.g., Poisson, normal, gamma). Within this family, the expected value of $Y_i$ is modeled as...
\[ g(\mu_i) = \mathbf{x}_i^\prime \mathbf{\beta} = \eta_i \]

for some function \(g(\cdot)\) (the link function), given a vector of \(\mathbf{x}_i\) of covariates and a vector of coefficients \(\mathbf{\beta}\).

The GLM is then composed of the three elements: a choice of distribution (e.g., binomial or gamma); a linear predictor \(\eta = \mathbf{x}_i^\prime \mathbf{\beta}\); a link function \(g\), such that \(E(Y) = \mu = g^{-1}(\eta)\).

Following Coe and Stern (1982), a two-stage approach has been used here to model rainfall occurrence (>1 mm month\(^{-1}\) rainfall) and amount separately. The given threshold of 1 mm, rather than 0 mm, to differentiate a wet from a dry month avoids some of the ambiguities that can be associated with the recording or estimation of very small rainfall amounts and is practically insignificant for agricultural and hydrological purposes.

The rainfall occurrence (pattern of dry–wet months) has been described using the following logistic regression:

\[ \ln \frac{p_i}{1 - p_i} = \mathbf{x}_i^\prime \mathbf{\beta}, \tag{1} \]

where \(p_i\) is the probability of rain for the \(i\)th case in the dataset conditional on a covariate vector \(\mathbf{x}_i\) with coefficient vector \(\mathbf{\beta}\).

To model the rainfall amounts during wet months, gamma distributions have been used. The rainfall amount for \(i\)th wet month has, conditional on a covariate vector \(\mathbf{\xi}_i\) and coefficient vector \(\mathbf{\gamma}\), a gamma distribution with mean \(\mu_i\), where

\[ \ln(\mu_i) = \mathbf{\xi}_i^\prime \mathbf{\gamma}. \tag{2} \]

The shape parameter of the gamma distribution, say \(\nu\), is assumed constant for all observations.

The gamma distribution has been widely used to represent monthly precipitation and has shown to be a good choice for water resources application in developing countries, because it is flexible enough to represent a variety of rainfall regimes and fairly familiar to the scientific community (Husak et al. 2007).

After choosing an appropriate set of covariates, the parameter vectors \(\mathbf{\beta}\) and \(\mathbf{\gamma}\) were estimated using maximum likelihood using the glm(·) routine in R. This assumes that the monthly rainfalls are conditionally independent given the covariates, whereas this is unlikely to be the case. The analysis therefore needs to account for temporal and spatial dependence as discussed later.

b. Model checking

To check for unexplained structure, mean Pearson residuals were computed for different subsets of observations, separately for the occurrence and the amount models. For an observation \(Y_i\), the Pearson residual is defined as follows (Chandler 2005):

\[ r_i = \frac{Y_i - \mu_i}{\sigma_i}, \]

where \(Y_i\) is the observed response for the \(i\)th case, \(\mu_i\) is the modeled mean, and \(\sigma_i\) is the standard deviation. If the fitted model is correct, then the Pearson residuals should all come from distributions with mean 0 and variance 1; hence, the standardized mean from any subset of \(n\) independent observation: say, \(\bar{r} = \sum r_i / \sqrt{n}\), should itself have mean 0 and variance 1. In addition, if \(n\) is large, then the standardized mean will have approximately a normal distribution. Thus, if standardized mean Pearson residuals are calculated for many different subsets (e.g., for each grid cell separately for each month of the year) and if the model is adequate, then around 95% of them should lie between \(-2\) and \(+2\).

c. Spatial and seasonal rainfall structure

To describe the climatology of the region, covariates representing regional and seasonal variation have been included in the models. Systematic regional variation is represented using Legendre polynomial transformations of latitude, longitude, and altitude (Abramowitz and Stegun 1965), whereas seasonality is described using a Fourier basis representing the annual cycle and two harmonics. Despite the use of 4 degrees of polynomials to represent regionality in the amount model, initial modeling attempts revealed that they were not sufficiently flexible to represent the effect of the Namib Desert, which is a relatively local but important feature. Therefore, a “Namib” variable—defined at each grid cell to take values 0, 0.3, 0.6, or 1, depending on the proximity to the Namib Desert—has been added to the model. Grid cells within the Namib are coded as 1, and those that are uninfluenced by the desert are coded as 0.

For the models to capture the persistence of periods of below/above-normal precipitation, terms representing the previous month’s rainfall have been included. For the occurrence model, the corresponding covariate is a binary (0 or 1) variable indicating whether rainfall occurred in the previous month at the same location. In the amount model, this covariate is replaced by \(\log(1 + \text{previous month’s rainfall})\).

d. Modeling strategy

Following Yan et al. (2002), the modeling process started with the development of a baseline model composed of basic factors influencing rainfall variability, such as geographical and seasonal factors. The progressive addition of the terms followed a perceived
order of importance, and insignificant factors were deleted throughout the procedure to keep the model manageable. The result of this initial stage was a model describing seasonal and regional variations in southern African rainfall variability. Pearson residual means were used to check the baseline structure, which was later improved. Subsequently, time-varying climate factors, which have been reported as influencing rainfall variability over the subcontinent, as summarized in section 2, were added. Initially, the statistical significance of each factor was examined individually and the most significant of the candidates was added to the model. The procedure was then repeated with the remaining factors, until no remaining factors were significant at the 1% level. Statistical significance was assessed using likelihood ratio tests, adjusted for intersite correlation using the methodology described in Chandler and Bate (2007). The 1% threshold for determining significance is more stringent than is commonly applied in climatological studies. This has been done in an attempt to avoid overfitting, since the dataset is so large that even small effects that are practically negligible (and hence of little interest) may appear statistically significant at more conventional levels.

Interactions, representing the possibility that the effect of one covariate depends on the values of others, were also considered. This allowed the models to represent, for example, differences in the phase of the seasonal cycle around the large area of study, as well as seasonal and regional variations in the strength of dependence between successive wet months, and in the effects of any climate covariate. In the procedure outlined above, the significance of each index was tested by adding the main effect along with seasonal and regional interactions, to allow for flexibility in representing the structures of those effects.

5. Baseline models

a. Occurrence model

In the baseline occurrence model, Legendre polynomials of degree two for altitude and degree three for latitude and longitude were used to describe the variation of precipitation occurrence with altitude and spatial locations, respectively. Seasonality and autocorrelation were represented using a Fourier basis and binary covariates, respectively, as described in section 4c. Besides the individual factors, the baseline model also included two- and three-way interactions between regional, seasonal, and autocorrelation terms. The interpretation of the three-way interactions is that the seasonal variation in autocorrelation itself varies over the subcontinent.

Standardized mean Pearson residuals (not shown) were computed for each month of the year at each grid cell. These indicate that the wet season (October–May) is extremely well represented throughout the subcontinent. During the remaining four months, in the dry season, from June to September, large standardized mean residuals occur mainly in the central part of the subcontinent, over Angola, Namibia, Botswana, Mozambique, and central South Africa. These are the areas where the model, which aims primarily to describe large-scale structures, is unable to capture relatively localized features characterizing the transition between highland and plain, desert and savannah. However, this is not considered too problematic, considering that the overall aim of the current research is the study of periods of water stress and drought for impact studies: in this context, it is more important to capture the characteristics of the rainy season than the dry, since it is here that a shortage of water may lead to potential disastrous impacts.

b. Amount model

The basic covariates in the best-fitting amount model are similar to those for the occurrence model. Some additions have been made, such as degree four Legendre polynomials for latitude and longitude, a degree three Legendre polynomial for altitude, and a term representing the Namib Desert local drying effect. In addition, autocorrelation is represented via a different transformation of the previous month’s rainfall, as described in section 4c.

Once again, the residuals indicate that the regional structure is well captured during the wet season (from October to April). Conversely, in the remaining 5 months, several grid cells have mean residuals that are significantly different from 0. Upon further investigation, however, these grid cells correspond to regions with very few wet months over the period used to fit the model. In particular, the highest residual is found in northern Botswana in June, where only 2 out of the 50 months experienced rainfall. In this case, since the sample size is so small the normal approximation, which underlies the classification of a standardized mean residual outside the range (−2, 2) as “problematic,” is likely to be extremely inaccurate. Moreover, in all cases with large standardized mean residuals, the maximum recorded rainfall was below 20 mm and the predicted amount below 2 mm. Once again, therefore, the poor performance is relatively unimportant in practical terms.

6. Final models

Having developed a baseline model, the potential effects of various climate indices were explored. Candidate
indices were initially identified from the literature, as reported in section 2. These candidate indices included atmospheric and oceanic fields that are physically meaningful and affect southern African rainfall variability, such as sea surface temperatures, sea level pressures, and relative humidity. For some fields, existing indices were considered (e.g., Niño-3.4, NAO, SAM; see Table 1). For other variables (e.g., relative humidity at 850 hPa over the region), to precisely select the most appropriate index, deeper analyses were also performed, using principal component analysis (PCA) over different regions, to identify dominant modes of variability related to precipitation variability and to reduce spatial fields to single predictors.

a. Occurrence model and effects of climate covariates

From the candidate external factors were selected, in order of significance:

1) The first principal component of the relative humidity field over southern Africa. The loading pattern of the index is characterized by a circular feature over the central area of southern Africa. High values of the index correspond to increased differences between the central area, driving a reduction in relative humidity; versus the surrounding areas, where increased humidity is experienced. Given the close link between precipitation and atmospheric moisture, it is not surprising that this index was the most statistically significant of those considered.

2) El Niño, considered the most important among the teleconnections affecting southern African rainfall and widely used to study its variability. The index used is the time series of the sea surface temperatures in the Niño-3.4 region.

3) SHT index, as representative of the warming trend.

4) Two variables representing the spatially and temporally varying effects of the subtropical jet stream. The first is $\text{INT}_t$, defined in section 3. The second is defined, for a location at latitude $\ell$ at time $t$, as

$$\text{POSEFF}_{\ell,t} = \max(0, \text{POS}_t - \ell),$$
which takes the value zero for locations to the north of the jet stream position at time $t$ and increases linearly with distance for locations to the south. As such, it can be regarded as a measure of the extent to which latitude $\ell$ is affected by midlatitude weather systems at time $t$. The motivation for this definition is deferred until the discussion of the amounts model below.

Having included all of these indices in the model, none of the remaining indexes in Table 1 was significant at the 1% level. The final model included 139 terms, of which 23 are main effects, 91 are two-way interactions, and 24 are three-way interactions. Such a large number of terms may seem, at first sight, excessive; however, the model was developed to represent seasonal and temporal rainfall variability over a large and topographically complex region. Indeed, 103 of the 139 terms relate to the nonparametric representation of regional variation. As Yan et al. (2006) pointed out, a model of this size for the representation of almost 50,000 data points is parsimonious if compared to a conventional approach of working separately with monthly anomalies at each grid cell (72 in this case).

To visualize the modeled effects of the various climate indices, we considered their contributions to the linear predictor in Eq. (1). Specifically, for each grid cell and each month we initially calculate the value of this linear predictor with all other covariates set to their average values. We then increase the value of each index in turn by one unit and recalculate the linear predictor. The change in the linear predictor represents the effect of a one-unit increase of the index upon the log odds of rainfall. Note that, because of interactions involving seasonal and regional covariates, this effect itself varies seasonally and regionally.

1) RELATIVE HUMIDITY INDEX

Figure 2 shows the effect on precipitation occurrence of a one-unit increase in the relative humidity index, defined by the first loading pattern over the subcontinent. This shows that an increase in the derived index is associated with a reduction in rainfall occurrence over the eastern and central areas and a corresponding increase elsewhere, particularly during the austral winter. Recall
that, because the way the index has been defined, an increase in the same corresponds to a reduction in relative humidity. The pattern found here is consistent with the first principal component loading pattern described above, although shifted slightly eastward.

2) **EL NINO INDEX**

El Niño occurrences (warm anomalies in the eastern Pacific) are associated with reduced rainfall in a large part of the subcontinent. However, El Niño events are not always accompanied by rainfall deficits (e.g., a strong El Niño occurred in 1997/98 without any coincident drier-than-normal conditions experienced in southern Africa). Figure 3 shows that according to the fitted occurrence model, an increase in the Niño index is associated with a reduction in the probability of rainfall occurrence during the austral summer over much of the subcontinent, especially over the southeastern part of it. A clear southeast–northwest gradient is also visible, with an increase in occurrence in the northwest, particularly from July to December. By negatively affecting the probability of rainfall over the southeast of the subcontinent, El Niño has a shortening effect upon the length of the rainy season. El Niño can therefore be summarized as reducing the amplitude of the seasonal cycle in rainfall occurrence in the southeast of southern Africa, mostly associated with a reduction in the probability of rainfall during the wet season in accordance with previous studies (see section 2).

3) **SHT INDEX**

Since the ocean tends to filter out high-frequency variation, SHT is a good index of underlying changes in global temperature. Its effect can therefore be interpreted substantially as a global warming signal. Figure 4 shows the modeled effect of a one-unit increase in SHT, which, in broad terms, corresponds to a reduction in the probability of precipitation occurrence throughout the year, except for the winter southern cape. This is consistent with the conclusions of Bates et al. (2008): higher SHTs are very likely to lead to an increase in the proportion of land surface in drought at any one time. In addition, a tendency for future drying in continental interiors during summer, especially in the subtropics, low and midlatitudes has been projected; and according to
Shongwe et al. (2009) with increasing temperature, the pattern of summer precipitation may be reduced, mainly visible in a later onset, and result in an eastward extension of desert areas in southern Africa. Again, Fig. 4 supports this because the modeled decreases in precipitation occurrence are greater in the east.

4) **JET STREAM EFFECTS**

Because of its complexity, the modeled effect of a strengthening and displacement of the subtropical jet stream cannot be easily shown with maps similar to the previous ones. Broadly speaking, inspection of the model coefficients suggests that a stronger jet stream intensity is associated with an increase in rainfall occurrence, particularly in locations south of the current jet stream position [see section 6b(6)].

**b. Amount model and effects of climate covariates**

In addition to the basic structure, the final amount model contains terms representing the effect of El Niño, SHT, relative humidity, and jet stream intensity and position as for the occurrence final model. In addition, two further climate covariates were found to be significant here as follows:

a) North–south gradient in southwest Indian Ocean sea surface temperatures. Standardized differences between the northwest (−25°–5°N, 60°–85°E) Indian Ocean SST standardized anomalies and southwest (−40°–−25°N, 40°–70°E) Indian Ocean SST standardized anomalies were computed to produce the index (anomalies computed with respect to the period 1961–90).

b) The SAM. The index used here is a mean sea level pressure (MSLP) difference between station data at both 40° and 65°S (see Table 1).

Geographical and seasonal interactions with all external factors were retained, as well as interactions with the autoregressive term for all but the Southern Hemisphere temperature.

More interactions were found to be significant than for the occurrence model, bringing the total number of

---

**Fig. 5.** As in Fig. 2, but for effect of relative humidity upon rainfall amount. Contours represent the multiplicative effect of a 1-unit increase in the relative humidity index on the expected monthly rainfall amount.
terms to 244 (33 main effects, and 186 two-way and 24 three-way interactions), of which 189 are related to the representation of regional variation and are needed to represent the complex topographic controls on rainfall in the area.

As for the occurrence model, the effect of each climate index can be visualized by considering the change in the linear predictor, resulting from a unit increase in the index. However, since a log link function has been used [Eq. (2)], a change of, say, $\delta$ in the linear predictor corresponds to multiplying the expected monthly rainfall by $\exp(\delta)$. We present the estimated multiplicative effects here, for ease in interpretation.

1) **RELATIVE HUMIDITY**

The pattern shown in Fig. 5 appears highly consistent with the loading pattern used as an index. The main feature is centrally situated over central southern Africa, in contrast with the effect modeled by the occurrence model, where the main effect was shifted toward the Indian Ocean (see Fig. 2). A one-unit increase in the index leads to a reduction in the expected monthly rainfall over the central area of the subcontinent, and to an increase around the edges.

2) **EL NIÑO**

As with the rainfall occurrence model, El Niño shows clear seasonally and locally varying effects (see Fig. 6). The figure shows the effect of a one-unit increase of the El Niño index conditional to the average values of the relative humidity index with which it interacts. The strongest impact is seen in the central and southern areas of southern Africa and in the austral summer: here, an increase in the Niño index is associated with a reduction in mean rainfall, in agreement with previous studies (Nicholson and Kim 1997).

3) **SOUTHERN HEMISPHERE MEAN TEMPERATURE**

As with the rainfall occurrence model, the modeled effect of an increase in SHT, conditional to the average values of the relative humidity index, results in an overall reduction in expected precipitation throughout the year.

![Figure 6](http://journals.ametsoc.org/doi/abs/10.1175/2010JCLI3924.1?journalCode=jcli)
especially over Namibia and Botswana (see Fig. 7), which are already characterized by dry climates.

4) SOUTHERN ANNULAR MODE

Figure 8 shows the modeled effect of a unit increase in the SAM index, which corresponds to enhanced westerlies in the southern oceans with a large-scale transfer of momentum and energy in the form of heat and moisture from the tropical to midlatitudes. The effect agrees with the results of Gillett et al. (2006): the positive phase of the SAM is associated with an increase in precipitation intensity over most of South Africa.

5) SOUTHWEST INDIAN OCEAN SSTS

The effect of the southwest Indian Ocean dipole index is shown in Fig. 9. In Washington and Preston (2006), in an idealized SSTs experiment, an increased SST gradient between a warm anomaly centered at 32°S, 55°E and a cold anomaly centered at 12°S, 65°E was associated with extremes in observed southern African austral summer rainfall, exerting control over the atmospheric circulation favorable for extreme rainfall. Figure 9 shows a similar pattern for the summer: an increase in the north–south SST gradient is associated with an increase in the expected monthly rainfall during late-summer months (mainly from February to May). However, the opposite is true for the remaining months.

6) JET STREAM INTENSITY AND POSITION

As for the occurrence model, these are represented via the variables \( \text{POSEFF}_{\ell t} \) and \( \text{INT}_t \) defined earlier. The variable \( \text{POSEFF}_{\ell t} \) was defined by considering that the effect of \( \text{INT}_t \) would probably vary spatially depending on \( \text{POS}_t \) (see section 3) relative to the location of interest. To gain some insight into the nature of this spatial variation, residuals were computed from the amounts model containing all of the factors listed above. For each grid cell, the relative position of the jet stream (in degrees north of the gridcell center) was then computed at each time point. The residuals were then grouped according to the values of this relative position and, for each group in turn, were regressed upon the jet stream.
intensity. Figure 10 shows the resulting regression coefficients for each group. This shows clearly that the regression coefficients are almost zero for groups corresponding to negative relative jet stream positions (i.e., for groups where the jet stream is to the south of the location of interest) and that it increases roughly linearly for positive values of the relative position. This figure shows that the regression coefficient of INT can itself be represented roughly as a linear function of the variable POSEFF\textsubscript{t}. This structure can be incorporated directly in a GLM: it is represented by the interaction between INT\textsubscript{t} and POSEFF\textsubscript{t}. The definition of POSEFF\textsubscript{t} is intended to mimic this pattern. The effect found in the current study is consistent with what was expressed by Richard et al. (2001) (see section 2), where a northward displacement of the jet stream could be linked to a northward displacement of the tracks of cyclonic systems.

As an additional check for unexplained structure in the final amount model, annual mean Pearson residuals for the whole region were computed and are shown in Fig. 11. Although no systematic trend is visible from the plot, there are runs of predominantly positive (e.g., during the 1980s) and negative (during the 1990s) values. This suggests the presence of interdecadal variability in precipitation over southern Africa that is not associated with any of the covariates in the fitted model, or with any of the other covariates considered (i.e., AMO and PDO, dominant modes of decadal variability).

Notice that, according to the fitted models, several of the covariates identified in the literature (section 2; Table 1) do not have a significant effect on precipitation in the region. Thus, although there may be physical arguments for such links, the analysis here suggests that their effects are not relevant at a subcontinental scale, at least after other, more important, factors have been accounted for. We note that some of those factors appeared significant initially; this illustrates the importance of accounting simultaneously for all relevant factors in analyses of this type. To investigate the possibility that the effects of some potentially important climate covariates were masked by co-linearities with other covariates in the model, we computed variance inflation factors (Fox 2002, ...
p. 216) for each of the covariates considered. Such an index provides a measure of the increase in the variance of an estimated regression coefficient due to co-linearity. Values in excess of 5 or 10 are generally considered to indicate problematic co-linearity (Montgomery and Peck 1992; Quinn and Keough 2002): here, all of the values were less than 2, so we conclude that collinearity is not a major problem for the modeling reported here.

7. Conclusions

The current study aimed to characterize southern African rainfall variability and gain deeper knowledge of its relationships with other aspects of the climate system. As far as we are aware, this is the first study to examine the simultaneous effects of many climate indices upon precipitation in the region. Such a comprehensive investigation of the climate system provides an understanding of how climate indices act in combination, as well as how the effect of one factor may itself be modulated by others.

Out of almost 20 atmospheric and oceanic features reported in the literature as potentially significant, only a half-dozen proved to be key drivers in explaining part of the variability in southern African rainfall. The ability

Fig. 9. As in Fig. 2, but for effect of SWIO SSTs.

Fig. 10. Definition of jet stream position variable POSEFF; coefficients from regression of residuals of penultimate rainfall amounts model upon INT, plotted separately for different relative latitudinal jet stream positions.
to isolate key drivers is another advantage of studying the various climate indices simultaneously, since this provides the opportunity to discriminate between competing explanations of the observed relationships. This illustrates the potential for analyses of single drivers to yield an incomplete understanding of the climate system. The findings reported here should provide a useful means of narrowing down candidate climatic drivers in future investigations of rainfall variability in specific parts of southern Africa.

The modeled effects of the main climate indices upon precipitation occurrences and amounts were broadly in agreement with what is already known about the climatic drivers of southern Africa rainfall. The isolation of the jet stream effect in explaining part of the rainfall variability in the considered region is new, however.

Conversely, where a statistical correlation between rainfall variability and climate index was not detected here, the current study does not necessarily contradict the previous research reported in section 2, since most of the climate indices did indeed appear significant in our analysis when considered separately. Thus, our work shows that after accounting for the most significant climate indices, any additional effect of the remainder is negligible.

Because of the complex orography of the region, many model terms were required to represent the regional variability across the subcontinent. The main time-varying factors affecting precipitation upon the studied region were first, relative humidity, followed by El Niño and mean Southern Hemisphere temperature.

A reviewer has asked whether the relationships indicated in this research could be exploited in a seasonal forecasting context. To investigate this, root-mean-square errors (RMSEs) were computed for three versions of the amounts model described above. The first model contained no climate indices and led to an RMSE of 39.52 mm. The second model contained only the potentially predictable climate covariates (Niño, SWIO, and SHT); the RMSE dropped to 38.91 mm. The final model, containing all of the selected climate covariates, had an RMSE of 37.29 mm. The rather small reduction in RMSE due to predictable climate covariates suggests that the scope for seasonal forecasting using these teleconnections is limited.

Interestingly, diagnostics revealed an interdecadal variability in rainfall amounts that could not be accounted for by any of the drivers considered. To understand the source of this variability would be a fruitful topic for future research.

The relationships revealed by the analysis reported here could, in principle, be used to downscale outputs from general circulation models to scales suitable for impact studies aiming to assess the hydrological impacts of climate change (see, e.g., Fealy and Sweeney 2007; Maraun et al. 2010). Work in this direction is ongoing.

Acknowledgments. Research conducted by Chiara Ambrosino was supported by a UCL Environment Institute Ph.D. scholarship. NCEP reanalysis data were provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, from their Web site (www.esrl.noaa.gov/psd/). The main author is thankful to Bryson Bates for his helpful advice.

REFERENCES

Conway, D., A. Persechino, S. Ardoin-Bardin, H. Hamandawana, C. Dieulin, and G. Mahc, 2009: Rainfall and water resources...


