Vol. 59: 39–60, 2014
doi: 10.3354/cr01198

CLIMATE RESEARCH
Clim Res

Published February 4

Spatial interpolation of daily rainfall stochastic generation parameters over East Africa

Pierre Camberlin[1,*], Wilson Gitau[2], Pascal Oettli[3], Laban Ogallo[4], Benjamin Bois[1]

[1]Centre de Recherches de Climatologie (CRC), Biogéosciences, UMR 6282 CNRS/Université de Bourgogne, 6 Bd Gabriel, 21000 Dijon, France

[2]Department of Meteorology, University of Nairobi, PO Box 30197, 00100 Nairobi, Kenya

[3]Department of Earth and Planetary Science, Graduate School of Science, The University of Tokyo, Science Building #1, Room 810, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

[4]IGAD Climate Prediction and Applications Centre (ICPAC), PO Box 10304, 00100 Nairobi, Kenya

ABSTRACT: Downscaling seasonal rainfall predictions to daily time-scale, for crop yield simulation for instance, can be performed using stochastic generators (SGs). The spatial interpolation of the SG parameters is required to generate rainfall time-series at ungauged places. A methodology is defined which makes use of topography to interpolate these parameters, in a region with a rugged terrain covering Kenya and north-eastern Tanzania. A first-order Markov chain was used to model rainfall occurrence, and a gamma distribution was used to model amounts. The 2 parameters of the Markov models, p01 and p11, and the 2 parameters of the gamma distribution are computed at 121 stations. The Kolmogorov–Smirnov test for goodness-of-fit shows that 88% (99%) of the stations and months have their dry (wet) spell frequencies successfully reproduced by first-order Markov chains, and two-third of the stations have their daily amounts satisfactorily fitted by the gamma distribution. Local regression, using elevation as the predictor and weighting stations according to distance from the target pixel and to environmental variables, is used to interpolate the 4 SG parameters. Cross-validation indicates that distance-weighted regression provides good estimates, but the inclusion of topographical variables (aspect in particular) improves the results further. The final maps show a strong orographic control of both the Markov and gamma parameters. However, while elevation has an effect on rainfall occurrence, rainfall intensity is more strongly related to local slope aspect, with eastward to southeastward oriented foothills and coastlines displaying the highest gamma scale values. These results suggest that a statistical disaggregation of daily rainfall is improved by taking explicitly into account topography through its effect on the spatial distribution of SG parameters.

KEY WORDS: Kenya · Tanzania · Disaggregation · Interpolation · Topography · Markov chains · Gamma distribution

1. INTRODUCTION

Disaggregating seasonal rainfall amounts into daily rainfall values is a key step in assessing the response of crops or hydrological systems to both interannual rainfall variability and long-term rainfall changes. Crop growth strongly depends on the intra-seasonal distribution of the rains. However, seasonal predictions (reflecting the interannual variability) only provide information on the seasonal amount of rain, and future projections (reflecting global climate change) are equally unable to satisfactorily project rainfall patterns at daily time-scales (Baron et al. 2005, Ines & Hansen 2006). In order to provide input to crop and hydrological models, it is necessary to define a methodology to generate high-resolution (in

*Corresponding author: camber@u-bourgogne.fr

time) rainfall time-series fitting both the expected seasonal anomaly and the mean local characteristics of rainfall. These characteristics are space-dependant, especially in areas displaying contrasting topography. They include the duration of the rainy season, the organisation of rainfall events in wet and dry spells and the intensity of rainfall events. Such characteristics of the within-season distribution of rains can be assessed by analysing observed daily rainfall time-series at available stations. However, they have to be spatially interpolated if one is to estimate crop growth or local runoff at locations for which climate data is unavailable, or for mapping purposes.

Among the different methods used to generate high-resolution (e.g. daily) time-series, stochastic generation is one of the most widely used (e.g. Wilby et al. 1998, Boulanger et al. 2007, Robertson et al. 2007). Stochastic models of daily rainfall often consist of 2 parts: a first sub-model for the occurrence of wet and dry days, and a second sub-model for the generation of rainfall amount on wet days (e.g. Stern & Coe 1984, Srikanthan & McMahon 2001).

The rainfall occurrence may be modelled by Markov chains, which define the state of each day as dry or wet and specify the relationship between the state of a given day and the state(s) of the preceding day(s). First-order Markov chains only consider the immediately preceding day. At tropical sites, Jones & Thornton (1993) suggest that, as a rule of thumb, first-order Markov chain models are to be replaced by higher order models. However, Deni et al. (2009) and Jimoh & Webster (1996) found that for Malaysia and Nigeria respectively, the daily rainfall occurrence is still often very adequately described by the Markov chain of the first-order model. Srikanthan & McMahon (2001) also consider that a first-order Markov chain remains adequate for many regions.

Many different distributions are used for modelling rainfall amounts. They include the mixed exponential distribution (Smith & Schreiber 1974, Woolhiser & Roldan 1986), various transformations of the normal distribution (Bardossy & Plate 1992, Chapman 1998), and the gamma distribution (Richardson & Wright 1984, Stern & Coe 1984). The latter has 2 parameters: α (shape parameter) and β (scale parameter).

The regionalisation of daily rainfall model parameters has been addressed by several studies, as summarized by Srikanthan & McMahon (2001). Woolhiser & Roldan (1986), for a limited number of stations in South Dakota, found that sophisticated interpolation methods were not yielding better than a simple arithmetic average of nearby stations. Kittel et al. (2004) used the WGEN stochastic generator (SG) to disaggregate monthly rainfall records at ungauged locations across the United States. Their WGEN parameterizations were assigned to cells based on the nearest daily station. It is anticipated that this method may result in step-like changes in the parameters, or the use of somewhat unrealistic parameters in areas of contrasted topography.

Given the unavailability of daily weather data in some countries, Geng et al. (1986) proposed the use of information from other, more easily available climate variables such as the mean monthly rainfall, the probability of a wet day or the mean daily rainfall amounts, to estimate rainfall generator parameters. They found that transitional probabilities used in Markov chain models were linearly related to the fraction of wet days per month, while the β parameter in a gamma distribution is closely related to the amount of rain per wet day. Jones & Thornton (1999, 2000) devised a method to estimate the parameters of a third-order Markov rainfall model using monthly mean climate normals and geographical coordinates, separately for a series of 664 station clusters representing worldwide climate types. Semenov & Brooks (1999) also proposed a spatial interpolation method of their LARS-WG stochastic generator, based on a local interpolation of the parameters, re-scaled to account for the effect of elevation on mean monthly precipitation amounts. The resulting daily time-series subsequently served as input for crop simulations in the UK (Semenov 2007). Wilks (2008) used local weighted regressions for the spatial interpolation of weather generator parameters. Elevation was used as a covariate, and successive regression models were defined at each pixel of the interpolation grid, with a weighting of the surrounding stations at which the parameters were available. The weights were a function of distance. It was found that this method performed better than other conventional interpolation methods, although for precipitation occurrence, it often has a skill similar to a domain-wide regression using latitude, longitude and elevation as predictors.

Castellvi et al. (2004) used a 2-part stochastic generation model for rainfall over Argentina. At stations without daily rainfall time-series, they used monthly rainfall amounts and monthly rainfall frequency to compute the parameters of the daily amounts under a gamma distribution. They found that their model was not adequate at sites near the Andes, where the relationship between monthly rainfall amounts and rainfall frequency was different from that determined at other locations.

In terms of spatial interpolation, East Africa is also a challenging zone because of its complicated orography and its variety of rainfall regimes (Ogallo 1993, Camberlin et al. 2009, Gitau et al. 2013). For this region, a few studies have been carried out on fitting Markov chains or various statistical distributions to daily rainfall occurrence and amounts. Barron et al. (2003) used a first-order Markov chain model to analyze the probability of dry spell occurrence at 2 semi-arid locations in Kenya and Tanzania. Oteng'i & Ogallo (1988) noted that first-order Markov chain models were providing good fits in highland areas of Kenya, but that both Markov and geometric models were not performing well in the arid environments of northern Kenya. Ochola & Kerkides (2003) and Biamah et al. (2005) found that first-order Markov chain models were relatively adequate for simulating wet and dry spells in Western Kenya and the drier watershed of Iuni in East-central Kenya, respectively. A Markov model has also been used to simulate the longest dry and wet spells at a subhumid station and a semi-arid one in Kenya (Sharma 1996). Arnold & Elliot (1996) predicted wet and dry spell lengths in Uganda using the CLIGEN weather generator, with good results at lower elevation sites but a poor fit at high elevation, low rainfall intensity sites. The weather generator was based on second-order Markov-chain models, but no comparison against first-order chain models was made. Gitau et al. (2008) found that over Kenya, first-order Markov chain models described the occurrence of wet and dry spells relatively well, and that an exponential (gamma) distribution was adequate in describing daily rainfall amounts exceeding 1 mm (5 mm). Hutchinson (1990) fitted a gamma distribution to daily rainfall in southern Somalia. He found little seasonal and geographical variations of the gamma shape parameter but the area under study is much flatter and more uniform than Kenya and Tanzania.

On the whole, these studies showed that Markov chains are generally adequate to simulate the distribution of rainfall events in the region, although with better results in wet than in dry areas for first-order chain models. These studies were at a local scale, often for a small number of stations, and it remains unknown whether the findings can be extrapolated to unsampled areas, especially in complex environments. In addition, few studies are available which attempted to model rainfall intensity, and its regionalization is poorly known.

The first objective of this study is thus to examine the spatial patterns of stochastic rainfall generation (SRG) parameters over part of East Africa. This knowledge is important for simulating daily rainfall time-series, as required in agroclimatic and hydrological models. It is also hypothesized that this knowledge enables a better understanding of rainfall producing mechanisms in the region, through the statistical properties of temporal rainfall distribution and their spatial variations, as suggested by Hills (1974). To these ends, the parameters used in some of the most common 2-part SRG (2-parameter first-order Markov chains for rainfall occurrence, and 2-parameter gamma distribution for rainfall amount) are considered at 121 stations in Kenya and northern Tanzania. The role of topography in the spatial distribution of the monthly values of these parameters is examined. An interpolation method is next applied which enables us to produce maps for each parameter, making use of their relationships with topography. The resulting maps are meant for the derivation of stochastically generated daily rainfall time-series at any location within the region under study.

2. DATA

The rainfall network (Fig. 1) consists of 121 stations whose daily data were obtained from the Kenya and Tanzanian Meteorological Departments. The region under study is diversified, including semi-arid lowlands in the north and the east (mean annual rainfall <500 mm) and much wetter areas (1000 to 2000 mm) along the coast, in the western Highlands of Kenya and at more isolated places in other mountain ranges (Fig. 1). There are generally 2 rainy seasons, from October to December (short rains) and from March to May (long rains) (Fig. 2). The months of January and February are drier, although occasional heavy rainfall can occur. The longer dry season (June to September) is highly contrasted, with much of Kenya and Tanzania getting no rain at all, but the Western Kenya Highlands, some parts of central Rift Valley and the coastal belt still experiencing heavy showers (Fig. 2).

In the dataset unfortunately, both the number of years available and the periods of records strongly differ between most of the stations. No single period of study could be defined; otherwise a huge proportion of the stations would have been lost. A minimum number of 15 yr, during the period 1961 to 2000, was considered for a station to be retained (i.e. 121 stations out of the initial data set of 180 stations). Most stations have data covering the years between 1961 and 1985, with 74% of the stations having less than 5 missing years during this period of time. The final

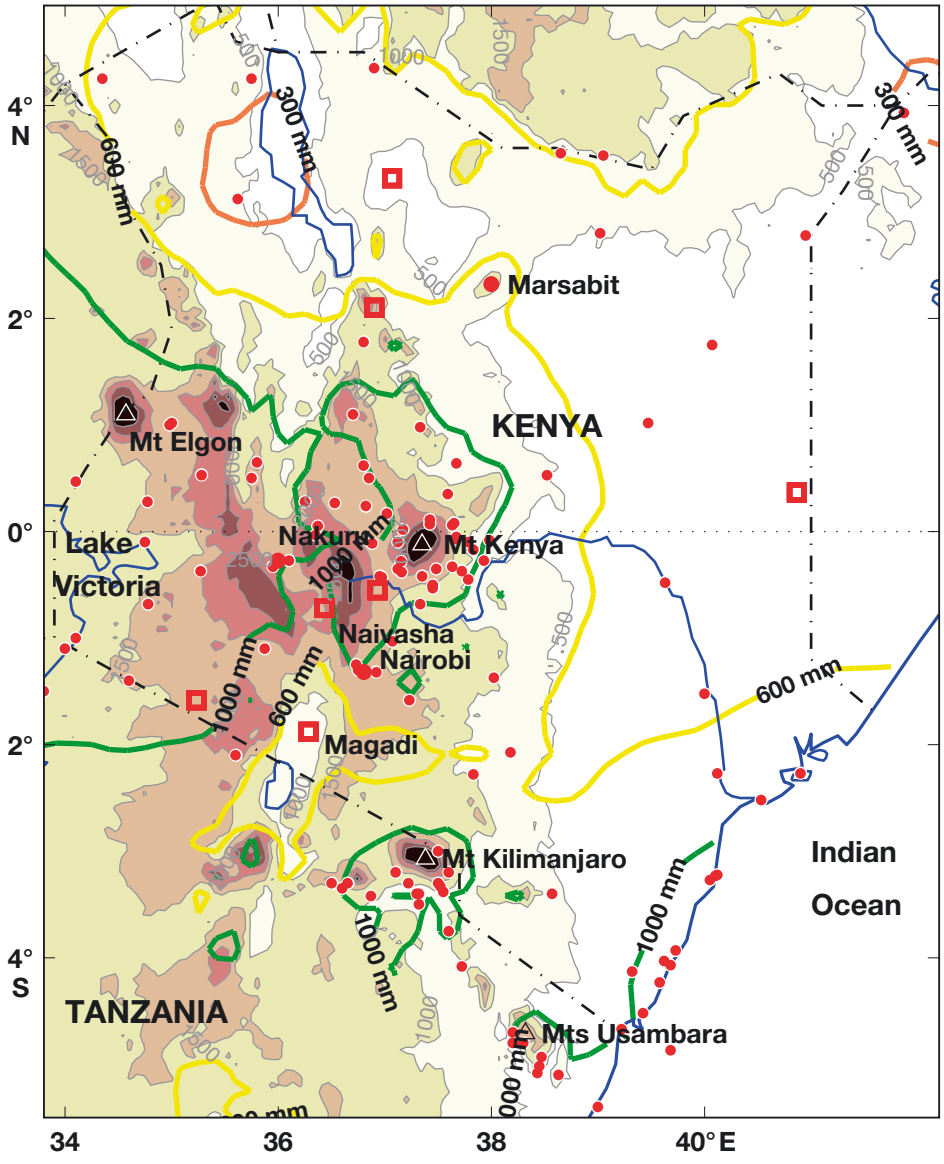


Fig. 1. Network of stations (red dots), terrain elevation (shadings and light grey contours, in m above sea level) and mean annual rainfall isohyets of 300 (orange thick contour), 600 (yellow) and 1000 mm (green). Red squares: 7 independent stations used in the validation only

network of 121 stations is not very dense but sufficient to sample reasonably well the different topographical settings. In very dry months, the parameters of the stochastic models (especially for the gamma distribution) sometimes show large and quite arbitrary variations. Therefore the stations and months recording a long-term average monthly rainfall <3 mm have been discarded for the interpolation step.

Using daily data raises the issue of data quality. While most of the stations are synoptic ones and provide good-quality data, this is more uneven for the others. Automatic quality control has been performed to detect and remove outliers, but there may remain a few inconsistencies in the data. Since falls <1 mm are irregularly reported at some stations, rain days in this study are defined as receiving at least 1 mm. Raingauges not monitored by professional observers may also suffer from the fact that rainfall amounts for 2 consecutive days (or more) are occasionally reported as a single (large) daily fall. Such cases are not easy to detect. However, a close analysis of day-by-day rainfall records failed to identify any such systematic bias in the data of the stations selected for further analysis.

To describe the topographical environment of each location, the Shuttle Radar Topography Mission

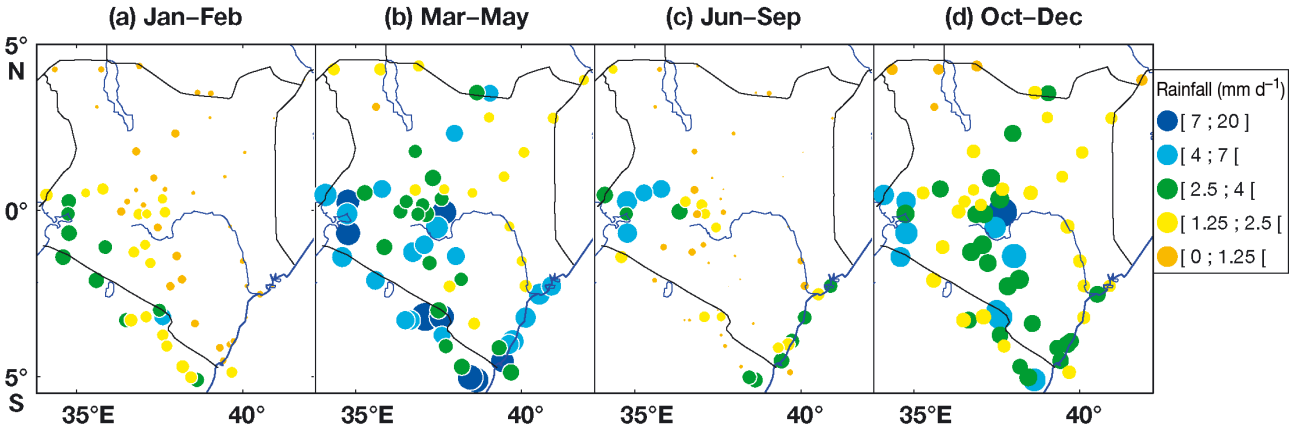


Fig. 2. Mean seasonal rainfall at all stations of the network used in the study. Size of dots proportional to mean daily rainfall (within the ranges defined by colour)

(SRTM30) digital elevation model (DEM) at a resolution of ~1 km (30 s arc), distributed by the US Geological Survey (USGS; Becker et al. 2009a), was used.

3. METHODOLOGY

The analysis consists of 3 steps. (1) The computation of the parameters of the weather generators at each available station; (2) the spatial interpolation of these parameters, and (3) the validation process, which enables the final mapping.

3.1. Computation of the SG parameters

Given the complex rainfall regimes in the region, all parameters of the SRG have been defined at a monthly time-scale. For the sake of simplicity and easier comparison with other studies, the parameters retained are among those most often used in 2-part SRG. The rainfall occurrence is defined by first-order Markov chains. Although they may not be the best in some tropical environments given their poor ability to reproduce long dry spells, they are the most often used and simulate reasonably well the rainfall occurrence in several regions including East Africa (e.g. Sharma 1996, Lana & Burgueño 1998, Wilks 1999, Gitau et al. 2008). Jimoh & Webster (1996) found that there was no discernible difference between the performances of the first- and second-order models over Nigeria. First-order Markov chains use 2 parameters which are transition probabilities between wet and dry days. p01 is the probability that a wet day follows a dry day and p11 is the probability that a wet day follows another wet day (see Appendix for details).

The rainfall amount for each wet day is given by a gamma distribution. Although not always the most suitable distribution for daily rainfall, again it is very commonly used and its parameters are useful in describing rainfall properties and their spatial variation (Husak et al. 2007, Becker et al. 2009b). The gamma distribution has 2 parameters (see Appendix): shape (α) and scale (β), which were estimated at each station and for each month via the method of maximum likelihood.

The goodness-of-fit of both the Markov chain models (Lana & Burgueño 1998) and the gamma distribution (Larsen & Pense 1982) for each station and each month is assessed using Kolmogorov–Smirnov (KS) test statistics. The KS statistics (at 99% confidence levels) are first used to compare the cumulative distribution functions of the lengths of wet and dry spells in the observation and as simulated by the Markov chain models, and then to compare the cumulative distribution functions of daily rainfall amounts in the observation and as obtained from the gamma distribution.

The spatial pattern of each parameter based on available station data is next examined to assess possible relationships with simple climatic variables (e.g. mean monthly rainfall) or environmental variables (e.g. altitude, slopes). This paves the way for the subsequent spatial interpolation.

3.2. Spatial interpolation of the parameters

The methodology for interpolating the parameters is a modified version of the methods proposed by Wilks (2008) and Daly et al. (2008). Wilks (2008) presented a flexible method which can capture nonlinear parameter variations in space, and applied it to

the estimation of weather generator parameters in the northeastern US. Daly et al. (2008) interpolated mean monthly precipitation and temperature in the US using the PRISM algorithm (Daly et al. 1994). A common feature of these studies is the use of weighted local regressions, also called geographically weighted regression (Fotheringham et al. 2002, Joly et al. 2011). In Wilks (2008), weights were defined according to the horizontal distance between training-data locations and the point of interpolation. In the PRISM algorithm, Daly et al (2008) considered a climate–elevation regression for each grid cell, with stations entering the regression being assigned weights based not only on distance but on the physiographic similarity of the station to the grid cell as well. In an earlier study, Johnson et al (2000) also used the PRISM algorithm to interpolate the parameters of weather generators over a mountainous region of the Northwestern US.

The weighted local regressions method was retained in this study. The main advantage of this method is that it allows spatially variable relationships between the parameters to be interpolated and geographical features. In the weighting process however, we make use of specific topographical descriptors previously shown to adequately explain rainfall distribution in southern Kenya (Oettli & Camberlin 2005). While Wilks (2008) did not consider topographical variables other than elevation, our study also differs from that by Daly et al. (2008) in that we use a few idealized topographical patterns instead of facet grids. The application of this methodology to an equatorial region which is very diverse in terms of topography and climate is also a challenge. Besides the validation of a method, the aim is to have a better description and understanding of how rainfall events are organized in East Africa.

The outline of the methodology is presented in Fig. 3, with further statistical details in the Appendix. For each pixel of the interpolation grid and for each parameter to be interpolated, a simple linear regression is defined using all surrounding stations, with elevation as the only predictor. At local level, elevation is a key variable in the distribution of mean monthly rainfall amounts, as demonstrated for the US (Daly et al. 2008) and East Africa (Hession & Moore 2011). Each station observation entering the regression model is weighted (Fig. 3). The weight is a function of several geographical properties considered separately or combined.

The first of these properties is the horizontal distance between the station and the target pixel. The corresponding weighting factor, DIST (Table 1), is inversely related to this distance, using a biweight kernel similar to that described in Wilks (2008). A weight of zero is assigned to stations located at a distance to the pixel exceeding a predefined threshold D_{MAX}. At a distance lower than D_{MAX}, the weights gradually increase from 0 to 1 as the station comes closer to the target pixel.

In most cases the number of stations available within the initial D_{MAX} radius and used in the local

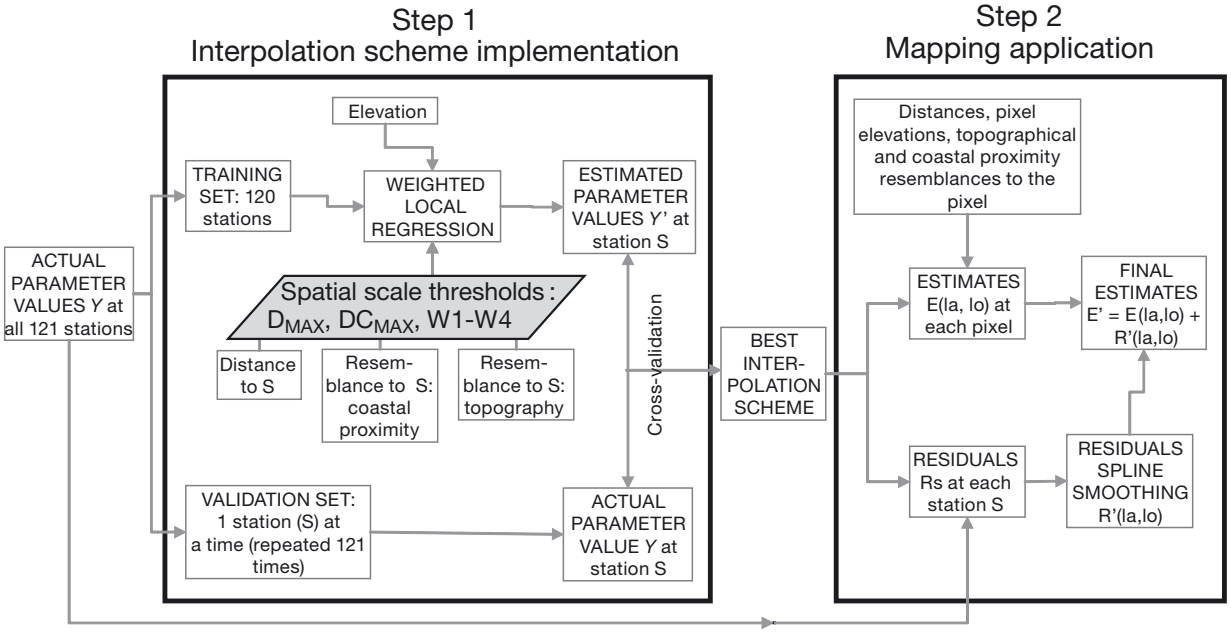


Fig. 3. Sketch-diagram of the methodology used for the interpolation of the stochastic generator parameters. la: latitude; lo: longitude; W1–W4: window size

linear regressions is >10, but a lower limit had to be set at 4 stations. In cases for which the number of 4 stations is not attained within the distance D_{MAX}, the value of D_{MAX} is gradually increased by steps of 0.5° until the required number of stations is reached. The minimum number of stations is set to a low value (4), but this aims at conveniently reflecting small-scale climate patterns in areas with a low density of stations (i.e. the northern and northeastern parts of the region), where a higher value would mean using quite distant stations for the interpolation.

Geographical properties other than DIST, not included in Wilks (2008) and partially only in Daly et al. (2008), have been added. They include the influence of water bodies and terrain (apart from elevation), which are important features governing rainfall distribution (in both space and time) in East Africa (Ogallo 1993, Song et al. 2004, Oettli & Camberlin 2005). One of these properties is the distance between the pixel and each station in terms of coastal proximity (COAST; Table 1). For the interpolation at subcoastal pixels, located at a maximum distance DC_{MAX} to the sea, maximum weight (1) is given to the stations which are located at the same distance to the sea as the target pixel. Weights linearly decrease for the other stations, as a function of the difference in coastal nearness between the target pixel and each station. For the pixels of the hinterland, where the influence of the sea should be negligible, all the stations apart from those of the coastal belt are given the same weight of 1. The stations of the coastal belt are then given a weight linearly related to the distance to the sea.

The last set of properties is related to the orographic settings. A series of orographic variables are defined, which are close to those previously used in a study dedicated to the interpolation of mean monthly rainfall fields over part of Tanzania and Kenya (Oettli & Camberlin 2005). These variables ($TOPO_1$ to $TOPO_8$) are listed in the bottom part of Table 1. They describe whether the pixel is on a slope (then described by its aspect: $TOPO_{1\,\&\,2}$), on a ridge or in a valley ($TOPO_{3–6}$), on a summit or in a basin

Table 1. List of geographical variables used in the weighting (selection) of stations entering the linear regression models. *W* indicates the weighting coefficients attached to each variable. The thresholds (D_{MAX} and DC_{MAX}) used in the computation of *W* are obtained by means of cross validation training (see Section 4.3)

Variable name	Description	Geographical feature 1 and corresponding weighting (*W*)	Geographical feature 2 and corresponding weighting (*W*)	Scales
DIST	Distance between target pixel *j* and station *i*	Stations close to the pixel $DIST_{ij} \leq D_{MAX}$: $W = (1 - DIST_{ij}^2 / D_{MAX}^2)^2$	Remote stations $DIST_{ij} > D_{MAX}$: $W = 0$	Pixel level
COAST	Coastal / inland position, based on distance to the sea (DSEA)	Subcoastal pixels $DSEA_j \leq DC_{MAX}$: $W = \max(1 - \lvert DSEA_i - DSEA_j \rvert / DC_{MAX}, 0)$	Inland pixels $DSEA_j > DC_{MAX}$: $W = \min(DSEA_i / DC_{MAX}, 1)$	Pixel level
$TOPO_1$ (W aspect)	West- or east-facing slopes	West-facing $1 \geq TOPO_1 > 0$	East-facing $-1 \leq TOPO_1 < 0$	4 scales of windows: W1 = 9 km W2 = 39 km W3 = 123 km W4 = 213 km
$TOPO_2$ (S aspect)	South- or north-facing slopes	South-facing $1 \geq TOPO_2 > 0$	North-facing $-1 \leq TOPO_2 < 0$	
$TOPO_3$ (RidgeN/S) $TOPO_4$ (RidgeE/W) $TOPO_5$ (RidgeNW/SE) $TOPO_6$ (RidgeNE/SW)	Ridges or valleys of various orientations	Ridge of a given orientation $1 \geq TOPO_k > 0$	Valley of a given orientation $-1 \leq TOPO_k < 0$	
$TOPO_7$ (summit)	Summit or depression	Summit $1 \geq TOPO_7 > 0$	Depression $-1 \leq TOPO_7 < 0$	
$TOPO_8$ (slope)	Average slope	Slope steepness $TOPO_8 > 0$	Flat terrain $TOPO_8 = 0$	
$TOPO_k$	Any topographical variable 1 to 8	$W_k = 1 - DTOPO_{ijk} / \max(DTOPO_{jk})$ with $DTOPO_{ijk} = (\lvert TOPO_{k,i} - TOPO_{k,j} \rvert)^{1/2}$		
ALL TOPO	Topographical variables 1 to 8 combined	$W = \Pi\, W_k$ with W_k the weights for topographical variable *k*		

($TOPO_7$), and what the average slope of the nearby pixels is ($TOPO_8$). These variables, except for the slope, are computed as idealized patterns derived from a principal component analysis of relative elevation around each pixel. Although some of these variables are not fully independent of each other, they are used for diagnostic purposes only, and in the weighting process. All these variables are computed using 4 different window sizes around the pixel, ranging from 9 to 213 km. Variables $TOPO_1$ to $TOPO_7$ are normalized, with values comprised between 1 (denoting that the variable is fully representative of the local topographical setting) and −1 (denoting the opposite topographical pattern, e.g. a valley instead of a ridge, see Table 1). The topographical setting (as depicted by these variables) at the candidate station to enter the regression is then compared to that of the target pixel. The square-root transformed difference between the values of the topographical variable at the station and at the target pixel is computed and normalized to get weighting coefficients ranging from 0 (e.g. the target pixel is on an east-facing slope while the candidate station is on a west-facing slope) to 1 (i.e. perfectly similar topographical settings). Slopes are also incorporated in this set of orographic properties. They are computed for each window using the maximum downhill slope method. Weighting is similar to that of the other topographical properties, except that slope values are normalized between 0 and 1.

Only one topographical variable (the one showing the highest skill for the month in consideration) is first used in the weighting process. Then a full model including all the 8 topographical variables (ALL TOPO) is defined, with the weights given as the product of all the 8 weighting coefficients. The distance, coastal proximity and topographical patterns (the best one, or all combined) are also used together in the final regression, with the weights obtained as the product of all the weighting coefficients.

3.3. Validation and mapping

Cross-validation is carried out in order to assess the general performance of the models, the part played by the different weighting factors, and to optimally assign the values of the parameters used in the weighting process (D_{MAX}, DC_{MAX} and size of the windows used to define topographical features). This is done by masking out each station in turn and estimating the target variable (Markov chain parameters and gamma parameters) at this station (Fig. 3). The root mean square error (RMSE) is used to determine the best combinations of weighting coefficients for the study region as a whole. The procedure is followed for each month separately since the spatial scales and topographical features which best explain the distribution of rainfall may vary seasonally. The pattern correlation coefficient between the observed and estimated values of the parameters is also computed.

Finally, the mapping of the SRG parameters is performed for each month using the optimum distance, coastal proximity indices, and topographical variables selected in the cross-validation step (Fig. 3). The parameters are estimated at all pixels of the SRTM30 DEM, within the envelope of available stations. Areas >2800 m in altitude are masked out since no stations above this elevation are available in the rainfall data set. In order to describe localized effects unaccounted for by the weighting procedure, and to ensure that interpolated values at each station are equal to observed values, a spline smoothing of the residuals is applied (ordinary kriging was also tested, resulting in quite a similar skill). The smoothed residuals are added to the estimates. Note that the residuals reflect both real, local-scale climate features, and sampling errors related to the small amount (and imperfect quality) of rainfall records available. Scrutinising the residuals maps makes us believe that the climate information overcomes the sampling errors, which justifies the use of the interpolated residuals in order to improve the final estimates.

4. RESULTS

4.1. Stochastic model parameterization

The goodness-of-fit for the simulation of precipitation occurrence was assessed using the Kolmogorov–Smirnov (KS) statistics. KS statistics were computed to compare the simulated cumulative distribution function of dry spell lengths (and wet spell lengths) to that of the empirical cumulative distribution. The test was carried out separately for each station and each month. In some instances, the number of 1 d dry spells is underestimated, and that of 4 to 10 d dry spells is overestimated. Similar results were obtained by Ochola & Kerkides (2003) for Western Kenya. However, for all stations combined, 88% (99%) of the months display no statistically significant difference between the simulated and empirical distributions of the dry (wet) spell lengths (Table 2). Several cases of statistically non-significant fitting

Table 2. Goodness-of-fit for Markov chain modelling of rainfall occurrence and gamma distribution of rainfall amounts, expressed as the percentage of stations for which the model successfully reproduces the observed patterns as from the Kolmogorov–Smirnov statistics (99% statistical significance). Top 2 rows: reproduction of dry and wet spell lengths by first-order Markov chain models; last row: reproduction of daily rainfall amounts by the gamma distribution

	Long rains									Short rains		
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Rainfall occurrence, Markov chain models												
Dry spells	83	94	83	93	86	88	87	91	93	92	87	75
Wet spells	99	99	100	98	99	100	99	99	99	100	99	99
Rainfall amounts, gamma distribution												
	52	45	67	75	74	72	72	69	68	70	73	71

seem to result from the relatively small number of years available at some stations (not shown). On the whole the first-order Markov model is very satisfactory in reproducing rainfall occurrence, even if in some instances higher-order Markov chains would probably be better suited to reproduce very long dry spells.

On a monthly basis (Table 2), slightly lower scores are obtained in December, January and March for the dry spells. This may result from the smaller number of rain days during these months which are located within or on the margins of the northern winter dry season, and the fact that rain is often associated with organised disturbances during this period. When mapped, the scores do not exhibit any clear spatial pattern (not shown) although the goodness-of-fit often tends to be higher at highland stations, again partly because of better sampling at these wetter locations.

For the simulation of rainfall amounts, as described by the gamma parameters, the KS statistics were also computed (Table 2). On average, 67% of the stations are successfully fitted (at the 99% confidence level), with scores ranging from 45% in February to 75% in April. These results can be seen as disappointing, but it was found that most cases of poor fitting were associated with low rainfall months and stations, for which the sample of days to determine the empirical probability distribution functions (PDF) was small. Empirical PDFs in such cases are not smooth, but it was visually found that the gamma distribution was still successfully fitting the general distribution of the daily rainfall amounts. This is confirmed by the fact that in wet months and at stations with a large sample of days, the KS test indicates most of the time that the gamma fitting is adequate.

Fig. 4 illustrates the marked contrasts in daily rainfall distributions across the network, as exemplified by the 2 stations of Nakuru (Rift Valley, West-Central Kenya) and Marsabit (on a volcanic mountain in northern Kenya). In April at the peak of the Long Rains, these 2 stations have fairly similar numbers of rain days (15.7 and 13.1 d for Nakuru and Marsabit, respectively). However, the rain-day intensity and frequency distribution at the 2 stations are very different. Nakuru, with a 95th percentile as low as 24 mm, has much less heavy rain days than Marsabit (95th percentile of 63 mm). This is portrayed by a scale parameter which is 4 times lower at Nakuru (5.6 mm) than at Marsabit (23 mm). The shape parameter, by contrast, is twice as large at Nakuru (1.5) as compared to Marsabit (0.8). Although at Marsabit for medium intensity rainfall (10 to 25 mm) the fit by the gamma distribution is not perfect, on the whole the difference between the 2 stations is adequately expressed by the gamma fitted distribution. This example shows the usefulness of the

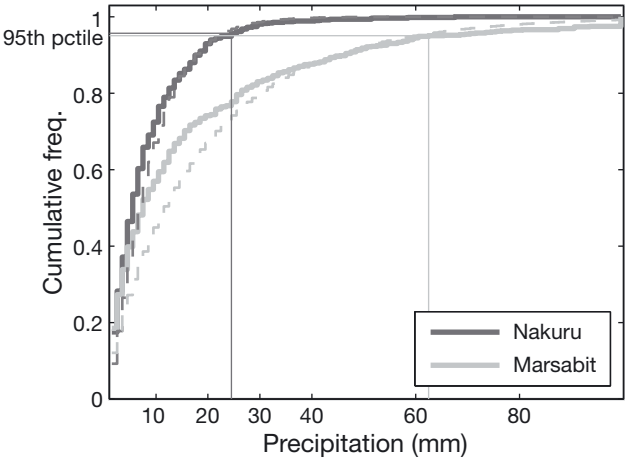


Fig. 4. Cumulative distribution function of April daily rainfall amounts for 2 stations with a similar number of rain days, but whose fitted gamma distributions have different shape and scale parameters: Nakuru (dark grey lines; scale parameter: 5.6 mm; shape parameter: 1.5) and Marsabit (light grey lines; scale parameter: 23.0 mm; shape parameter: 0.8). Solid bold lines: empirical distributions; dashed lines: fitted gamma distribution. Empirical 95th percentiles are also displayed

gamma parameters in describing key aspects of rainfall distribution, which strongly impact on runoff, erosion and crop water balance.

4.2. Spatial patterns of the parameters

4.2.1. Precipitation occurrence (Markov chains)

The spatial patterns of p01 (probability that a wet day follows a dry day) and p11 (probability that a wet day follows another wet day) are shown for the month of April as an example (Fig. 5). The 2 maps are relatively similar, with higher values concentrated in the highlands and Lake Region; and lower values in the eastern and northern plains. These patterns are quite similar to the distribution of mean monthly rainfall as from published maps (not shown, see for instance Said et al. 2007), an indication that rainfall occurrence is the main determinant in the mean spatial variations of the rains. Fig. 6 indicates that the p01 and p11 parameters are actually strongly related to mean monthly rainfall amounts in each month of the year. The lowest correlation value is found for p11 in January, but it remains highly significant. However, there are also a few distinct features. Within the highlands, some slopes showing a specific aspect, enjoying well-defined upslope breezes (slopes facing lake Victoria, southern slopes of the Kilimanjaro and Usambara ranges in Tanzania) tend to have higher values of both p01 and p11 (Fig. 5). The coast also stands out by having quite low p01 values (although its mean rainfall is high during this season) and higher p11 values. This is indicative that rain days tend to be associated with more organized disturbances, lasting a few days, than elsewhere in East Africa.

The domain-wide relationship between the Markov chain parameters and topography was examined. Elevation is significantly correlated (99% confidence level) with both parameters for every month except during the northern summer, between May and September (not shown). Other significant correlations are found with some topographical variables describing slope, aspect and terrain geometry. However, much of these relationships are related to the covariation between mean monthly rainfall and altitude. Partial correlations, controlling for the effect of mean

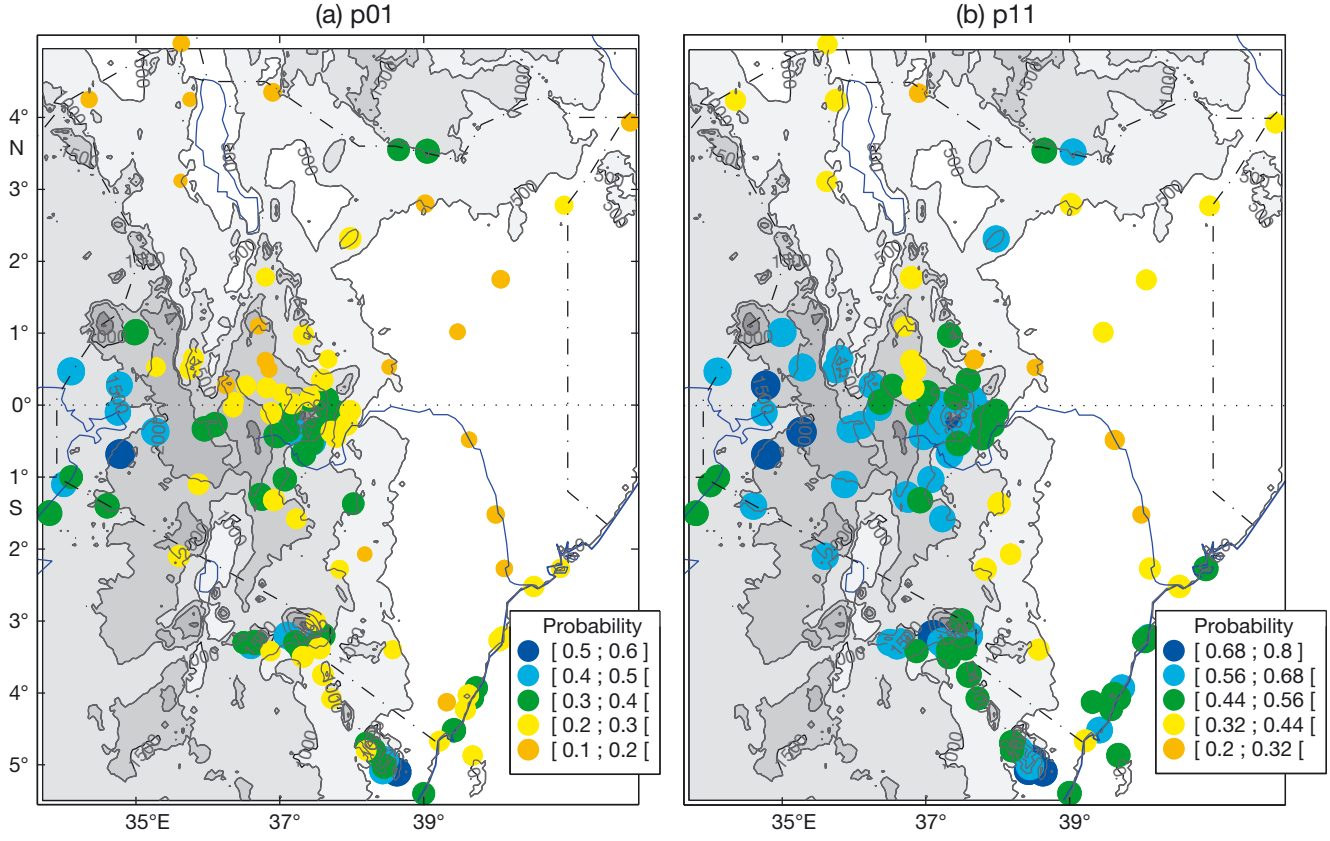


Fig. 5. Spatial patterns of the conditional probabilities of (a) a wet day following a dry day (p01), and (b) a wet day following another wet day (p11) for the month of April. Shadings: elevation (m). Size of dots proportional to mean daily rainfall (within the ranges defined by colour)

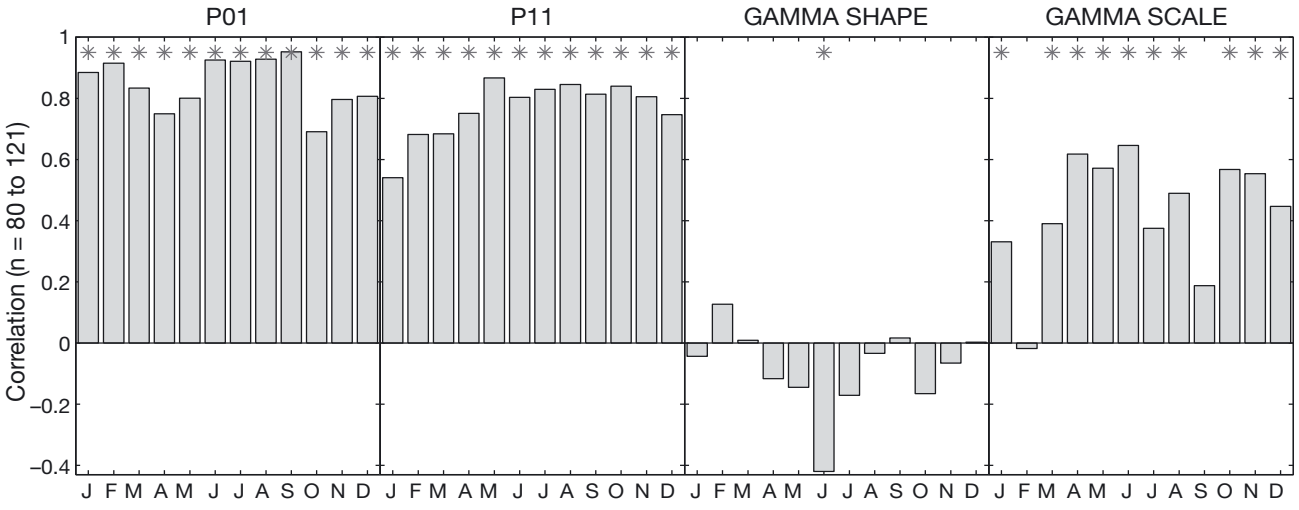


Fig. 6. Pattern correlations between mean monthly rainfall amounts and parameters of the stochastic generator: conditional probabilities of a wet day (p01 and p11), gamma shape and scale parameters. Asterisks: significant at $p < 0.05$

monthly rainfall, reveals that the remaining topographical signal in both p01 and p11 is small and significant only for a few variables (especially west/east aspect). Moreover, the strong relationship with mean monthly rainfall suggests that for the mapping of p01 and p11, an alternative would be to use already published interpolated or gridded mean rainfall maps. However, using linear functions to map these values could have problems in not bounding the transition probabilities on the unit interval. While there is an analytic relationship between the 2 first-order Markov transition probabilities and the average number of wet days (Wilks & Wilby 1999), there is no high-resolution map of the latter variable over East Africa, ruling out any use for a subsequent mapping of the p01 and p11 fields.

4.2.2. Precipitation amounts (gamma parameters)

The spatial patterns of the shape and scale parameters are shown for the month of April again as an example (Fig. 7). Both maps display a clear east–west contrast. Low shape parameter values are found in the east, from the coast to the eastern slopes, including highland stations such as those of Mt Kilimanjaro, and Marsabit in the north. High values are found in the west (especially in the highlands) and scattered locations in Tanzania. The scale parameter shows almost the opposite pattern, with high values in the east, low in the west. There seems to be systematically high values of the scale parameter at windward locations (facing east or south-east), which include the Indian Ocean coast, the southern slopes of Mt Kilimanjaro, the south-eastern slopes of Mt Kenya, Mt Marsabit, and several stations along the gentle slopes in east-central Kenya. Stations located on leeward slopes or in the western highlands have low values. These apparent relationships with topography will be examined more closely below. The dry stations of northern and eastern Kenya have intermediate values.

On the whole, these patterns are, as expected, much more weakly related to mean monthly rainfall amounts than the Markov chain parameters (Fig. 6). The shape parameter shows no significant correlations with mean monthly rainfall, except for a weak negative correlation in June. The scale parameter is positively correlated with mean monthly rainfall, but the correlation remains low and is even non significant in February and September. For Argentina (Castellvi et al. 2004), the scale and shape parameters have been shown to be strongly related to the mean rainfall intensity per rain-day (Geng et al. 1986). The product of the shape and scale parameters of the gamma function equals the mean daily wet-day amount. However, data on daily mean rainfall intensity and its spatial pattern are not widely available, thus precluding the use of this variable for the regionalisation of gamma parameters. This justifies the present analysis with regard to the possible role played by topography on the geography of the gamma shape and scale parameters.

The month-by-month correlations between the spatial patterns of the gamma parameters and topographical variables are shown in Figs. 8 & 9. The shape parameter (Fig. 8) is significantly and positively correlated (at 99% confidence levels) with the

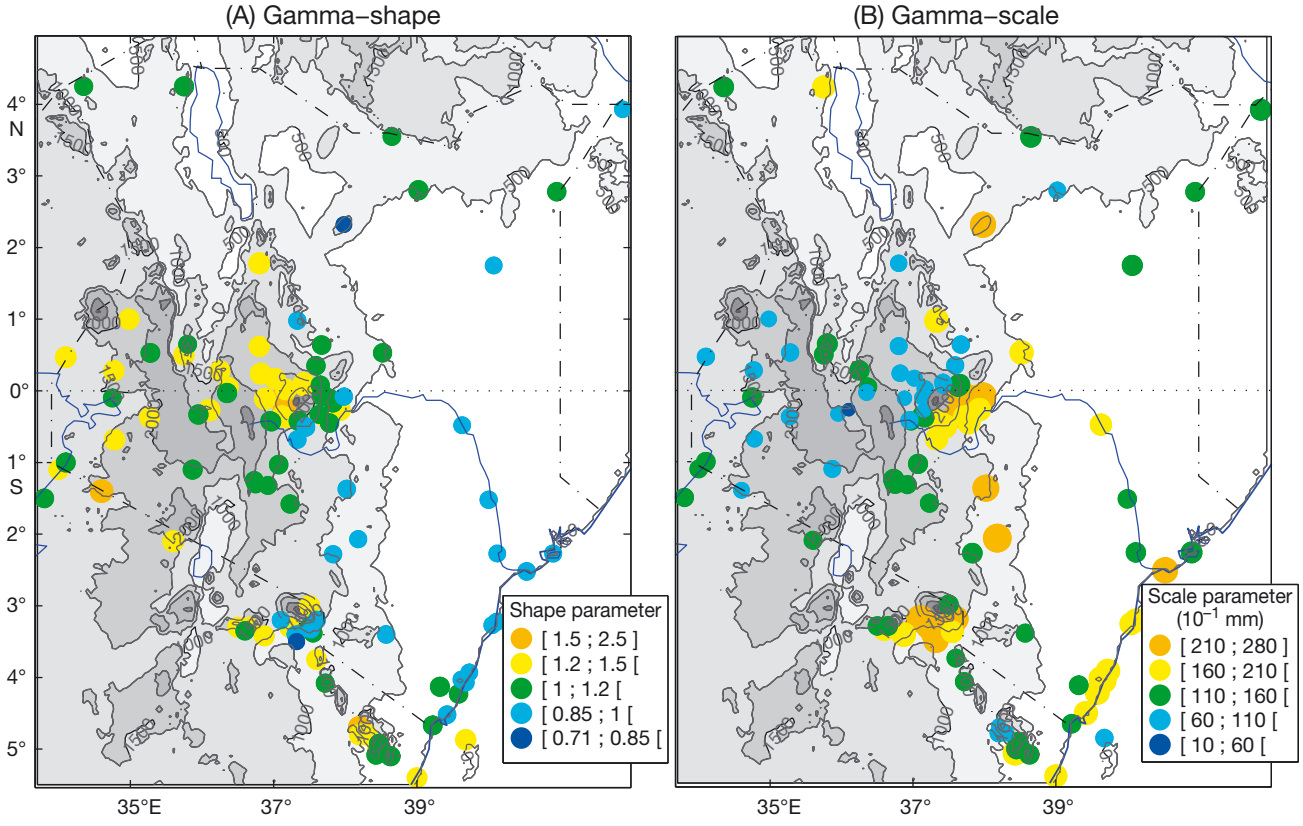


Fig. 7. Spatial patterns of the gamma (a) shape and (b) scale parameters for the month of April. Shape parameter is dimensionless. Size of dots proportional to mean daily rainfall (within the ranges defined by colour)

altitude of the stations in all months but July. The highest influence of altitude ($r > 0.5$) is found around the peak months of both rainy seasons (April, October–November). For the analysis of the relationships with other topographical variables, partial correlations are used which allow the removal of the effect of altitude, given that altitude is not linearly independent from some variables such as slope. The results are meant as guidelines only, since some collinearity remains between variables $TOPO_3$ to $TOPO_7$. The most significant correlations are for topographical variables defined at large scale, W4 (213 km). Higher shape values are found over steep highland areas in June to August. During the rainy seasons, it is mostly the west/east aspect which controls the shape values, with higher (lower) values at west (east)-facing locations. Large-scale ridges also slightly modulate the gamma shape parameter. These results suggest that during the rainy seasons, a greater proportion of heavy rainfall events (resulting in a less skewed distribution, and a lower shape parameter) are recorded at windward, east-facing stations, and at low elevations.

The scale parameter also is strongly controlled by the elevation of the stations (Fig. 9), although in this case the correlation is negative. The negative correlation reflects the fact that the product of shape and scale equals mean daily wet-day amount. Lower scale values are found in the highlands during most of the year. This denotes the lower mean daily rainfall intensities of the highlands, except for some windward areas (Nieuwolt 1974, Moore 1979). Other topographical variables significantly influence the value of the scale. During the Long Rains (March to May), a combination of south and east aspects (i.e. windward locations) tends to display higher scale values (Fig. 9, top right panels). In the dry northern summer season, peaking in July–August, a totally opposite pattern prevails, in which the south-east aspect gets very low scale values. It still corresponds to windward locations, but in a stable atmosphere which results in very low rainfall amounts. The October–November rainy season looks more alike the previous (March to May) rainy season, except that only $TOPO_1$, opposing east and west aspects, has a significant impact on the scale parameter. This is in line with the low-level winds being due east. During the rainy seasons, the effect of aspect is best shown at intermediate (39 or 123 km) window sizes. Summit locations also tend to have an

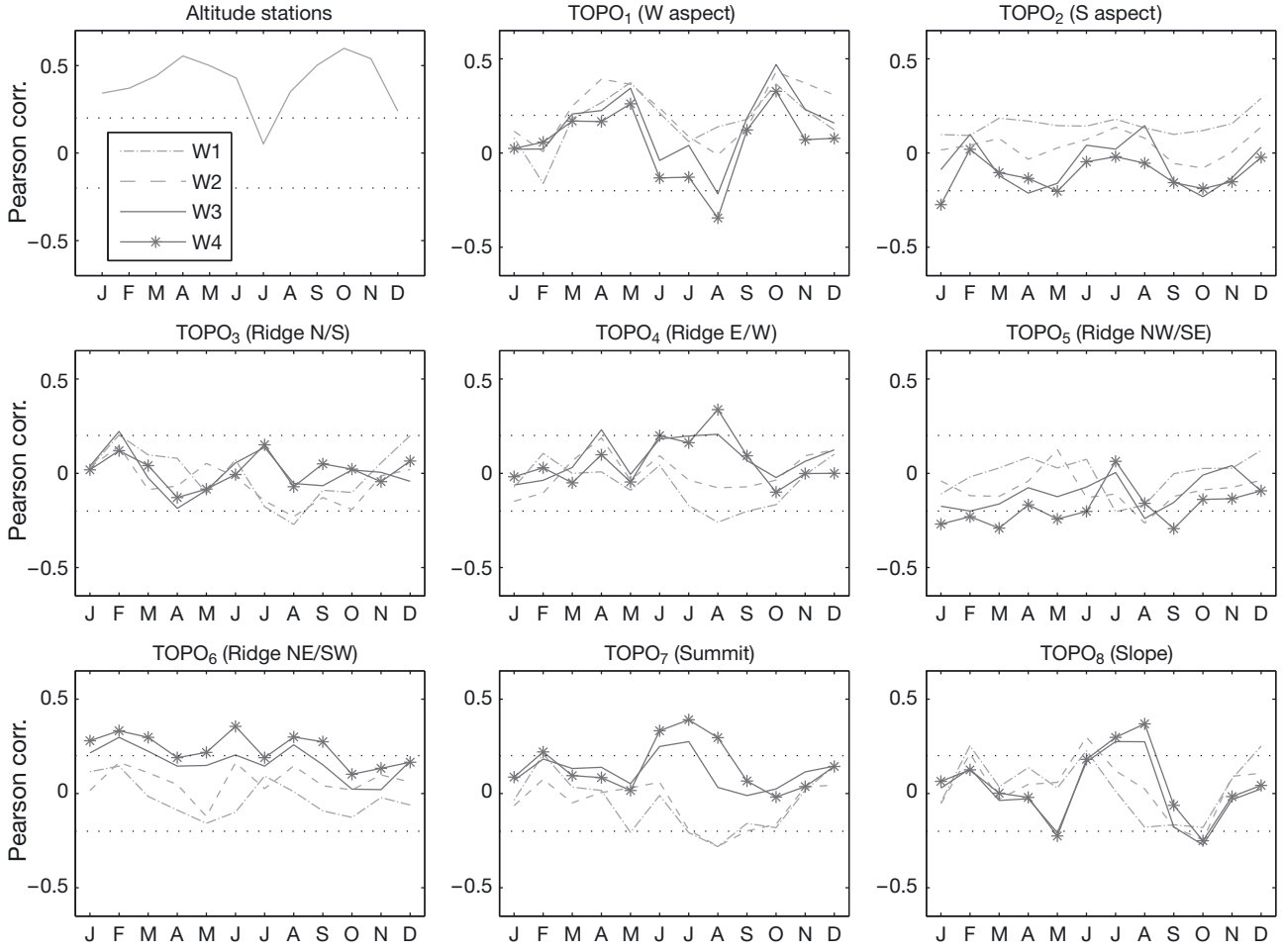


Fig. 8. Month-by-month correlations between the spatial patterns of the shape parameter of the gamma distribution and topographical variables. For all variables except the first one, partial correlations are used which remove the effect of altitude. W1, W2, W3 and W4: 4 window sizes used to compute the topographical variables (altitude is directly obtained from the SRTM DEM, thus no windowing is applied to this variable). Horizontal dotted lines: correlation values associated with 0.01 significance level

effect in the June to September season, though the correlations are of opposite signs at small and large scales, showing that different mechanisms jointly intervene in the spatial variations of daily rainfall intensity. Finally, slopes are also strongly correlated with the scale values, which increase along the steeper slopes during the rain seasons and decrease during the June to September period. On the whole, these results suggest that orographic uplift associated with the combination of terrain geometry and low level wind flow play a very important part in the spatial distribution of the scale parameter. While these relationships are global (domain-wide), we can expect an even stronger influence of topography at a local scale. In order to interpolate the SRG parameters, the next step considers local regressions with elevation, with a weighting based on environmental variables.

4.3. Spatial interpolation

The spatial interpolation first requires us to test the usefulness of the various environmental variables (topography as well as distance to the ocean) in the estimation of each parameter of the SGs. It also requires fixing some thresholds (D_{MAX}, DC_{MAX}) in the weighting of stations used. This is typically accomplished using cross-validation, where each of the 121 stations is left out in turn, and each of the 4 SRG parameters is estimated on a monthly basis. The performance of the various models, incorporating one or several weighting variables, is assessed using RMSE and the Pearson correlation coefficient, allowing us to test how close the interpolated values are from the observation. The interpolation errors (RMSE) are shown in Table 3 for 2 of the stochastic parameters (p01 and the gamma scale parameter), for the month

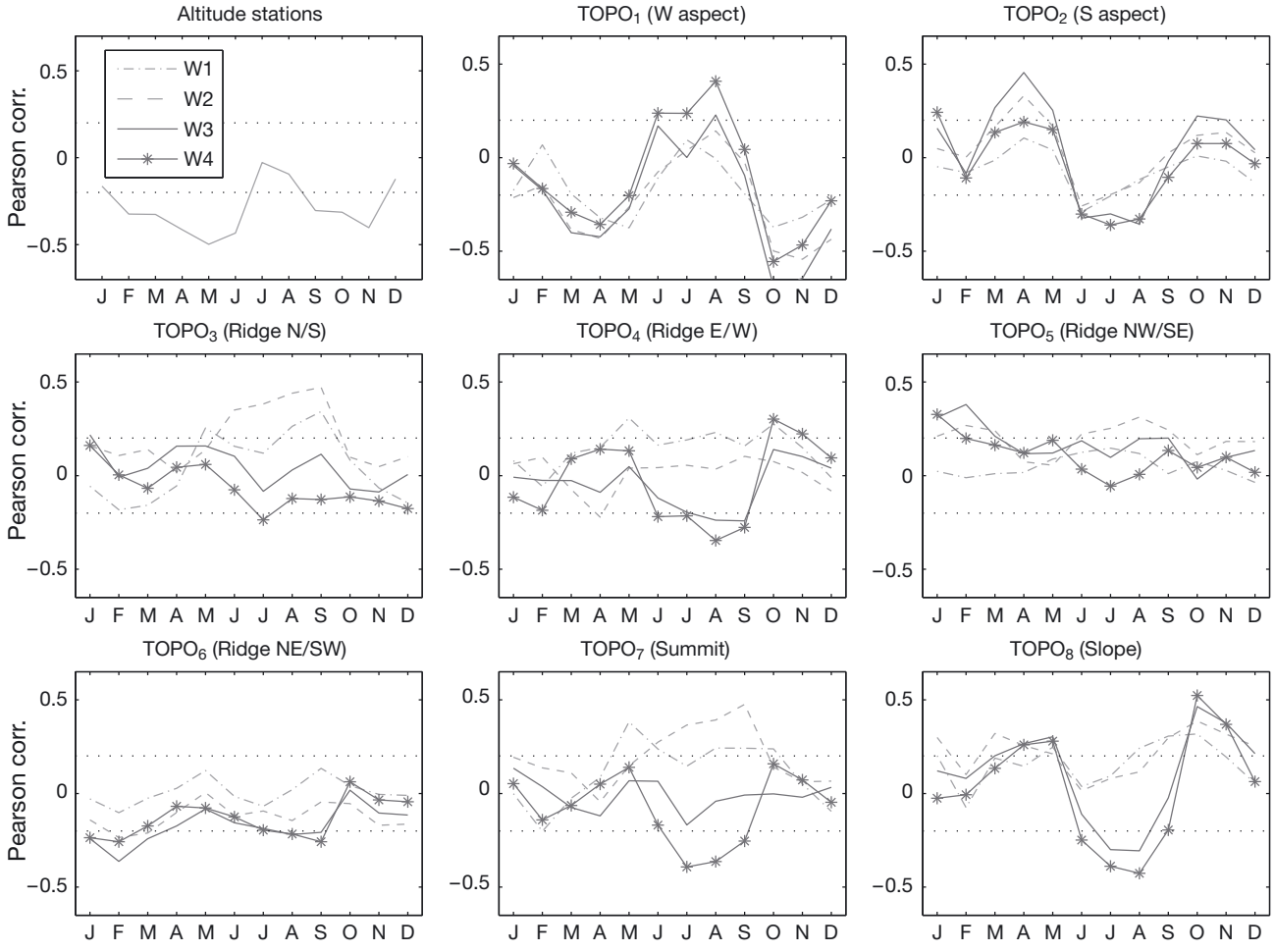


Fig. 9. Same as for Fig. 8 but for the scale parameter of the gamma distribution

of April taken as an example (the wettest month in East Africa as a whole), and as a summary for all months of the year. As a benchmark, we also define a model in which all the 120 stations (121 minus the hidden station) are given the same weight in the regression equation used to estimate the values at the hidden station (i.e. using a global—rather than a local—weighted regression with elevation).

Compared to the benchmark experiment, the use of a single environmental variable (either coastal proximity or topography) to weigh stations, without considering their distance to the station to be predicted, moderately reduces the RMSE of both p01 and the gamma scale parameter (Table 3). However, their combination results in a further RMSE decrease, especially for p01. The most relevant topographical variable is generally west/east or south/north aspect ($TOPO_1$ and $TOPO_2$, respectively), at the 123 km window size (W3), in agreement with the above findings. The exact value of the DSEA parameter (limit of ocean influence) has only a small effect, and it actually varies from month to month.

Weighting stations according to distance generally has a strong influence on skill. This allows the depiction of spatially varying relationships between rainfall and elevation. Except for gamma scale in some months, distance weighting generally outperforms coastal proximity and topography for the interpolation of the target variables (Table 3). Optimum D_{MAX} values are relatively small for the wettest months (50 to 150 km), but larger for the gamma scale parameter between May and September.

Finally, the combination of distance, coastal proximity and topography in the weighting process further improves the skill of the interpolation, as shown by the RMSE (Table 3). The lowest RMSE is generally obtained when all the topographical variables are combined in the weighting process. This justifies the use of all the environmental and distance variables as weighting factors in the final interpolation (below; this section).

Table 3. Best interpolation models of the p01 and gamma-scale parameters. Values are errors (RMSE, in cross-validation mode) for models incorporating different weighting variables. Parentheses: lowest RMSE-yielding variables (i.e. the best model). In the weighting of the stations for the interpolation, all weights except those relative to the tested variable(s) are set to 1

Variable(s) used in weighting	Rainfall parameters to be interpolated			
	p01		Gamma-scale (10^{-1} mm)	
	April	Average for all months	April	Average for all months
None (all weights = 1)	0.091	0.084	48.2	34.6
COAST	0.089 (DC_{MAX} = 150 km)	0.077 (DC_{MAX} = 150–300 km)	46.9 (DC_{MAX} = 300 km)	34.1 (DC_{MAX} = 50–300 km depending on month)
Best topographical variable	0.082 ($TOPO_2$ W3)	0.078 (May–Nov: $TOPO_1$, W4; Dec–Apr: $TOPO_2$, W4 or W3)	45.7 ($TOPO_2$ W3)	32.5 ($TOPO_1$ W3)
COAST + best topographical variable	0.079 (DC_{MAX} = 150 km, $TOPO_2$ W3)	0.07 (DC_{MAX} = 100 to 300 km, $TOPO_1$ or $TOPO_2$, W3 or W4)	44.2 (DC_{MAX} = 300 km, $TOPO_2$ W3)	31.9 (DC_{MAX} = 100 to 300 km, $TOPO_1$ or $TOPO_2$, W3)
DIST	0.074 (D_{MAX}= 150 km)	0.055 (D_{MAX} = 50 to 150 km)	41.6 (D_{MAX}= 50 km)	32.2 (D_{MAX} = 50 km in wet months, 500 km in May–Sep)
DIST + COAST + best topographical variable	0.067 (D_{MAX}= 200 km, DC_{MAX} = 200 km, $TOPO_2$ W3)	0.05 (varying combinations depending on month)	37.6 (D_{MAX} = 200 km, DC_{MAX} = 200 km, $TOPO_2$ W3)	29.8 (varying combinations depending on month)
DIST + COAST + ALL TOPO	0.064 (D_{MAX} = 200 km, DC_{MAX} = 200 km)	0.048 (varying combinations depending on month)	36.4 (D_{MAX} = 200 km, DC_{MAX} = 200 km)	29.3 (varying combinations depending on month)

Fig. 10 summarizes the results on a monthly basis, by showing the RMSE and pattern correlation coefficient between the observed and estimated values of p01 and gamma scale. For p01, the use of weighted local regressions strongly improves the estimation, as compared to the use of a global regression (all station weights set to 1). In particular, the pattern correlation is >0.80 in most months (Fig. 10b). The inclusion of topographical information in the weighting slightly improves the results as compared to the use of distance only, especially from April to December. The role of topography is more decisive for the gamma scale. Although correlations between observations and estimations are lower than for p01, considering topography significantly reduces the errors in wet season months (Fig. 10c) and enhances the correlations during most of the year (Fig. 10d). This justifies the use of the fully weighted models (including distance, coastal proximity and topographical variables) in the final interpolation.

The resulting map of the gamma scale parameter for the month of April, taken as an example, is shown in Fig. 11. It should be compared to the observed station data as displayed in Fig. 7b. High values are found in the foothills and to the east of the Kenya highlands, of Mt Kilimanjaro and Mt Meru in Tanzania, and along the coast. Although not necessarily steep, these locations have in common that they stand out as the first obstacles encountered by the moist air parcels advected from the Indian Ocean by the easterly to south-easterly low-level flow to East Africa (Camberlin et al. in press). These large gamma scale values denote that the proportion of heavy rainfall events at these locations is high. This could result from the triggering of moist convection, in an unstable atmosphere, by relatively minor topographical features. By contrast, low values are found at high elevations, especially on leeward slopes (i.e. facing west or north-west). This includes much of the Central and

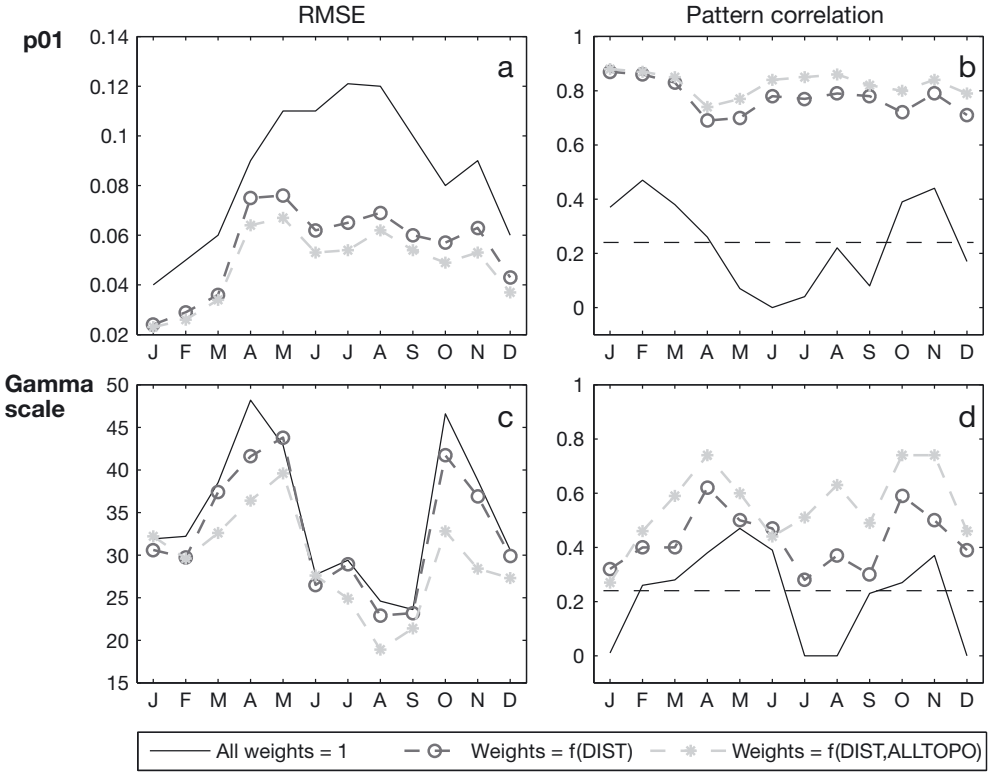
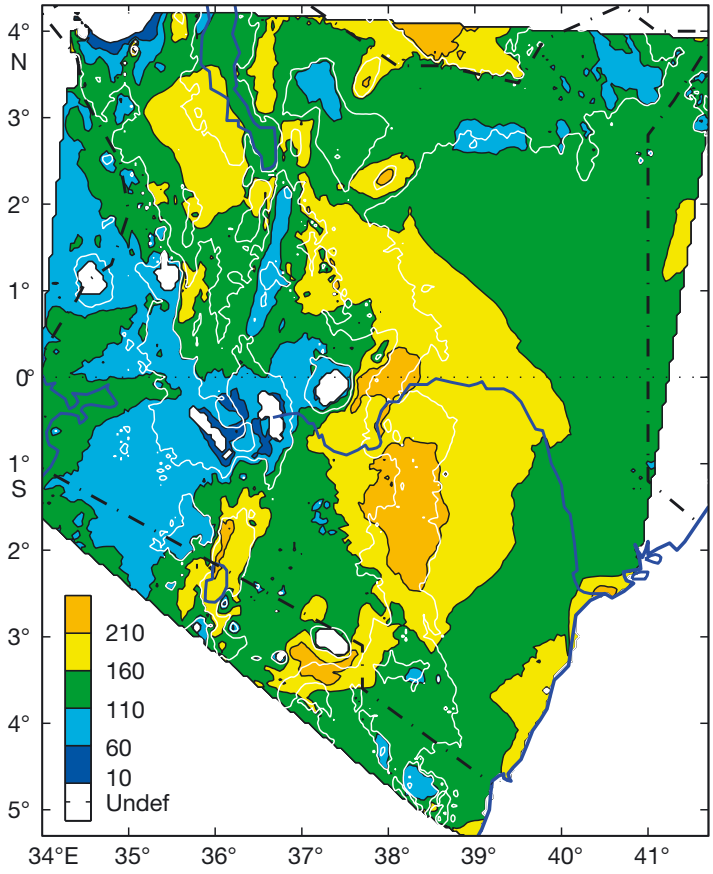


Fig. 10. Cross-validated skill scores of the estimated values of the stochastic generator parameters: (a,b) p01 and (c,d) gamma scale (*y*-axis unit in [c]: 10^{-1} mm). (a,c) RMSE values, (b,d) pattern correlation between the observed and estimated values (n = 121). Solid line: estimates based on a simple linear regression with altitude (no weighting); dashed line with circles: stations weighted according to distance; dashed line with stars: stations weighted according to distance and all topographical variables. Horizontal dashed lines (panels b,d): 0.01 significance level



Western Kenya Highlands, including the north-western slopes of Mt Kenya. More localised low values are found on the leeward (north-western) side of the Usambara Range, in northern Tanzania closer to the coast. They are indicative of lower daily rainfall intensities. This can be explained by the lower moisture content at high elevations and along leeward slopes, and the fact that orographic uplift is less efficient than along windward slopes.

The validity of the map is further tested using additional stations with relatively short records, which were not incorporated in the data base. A selection was made of 7 stations (Fig. 1: red squares), purposely quite distant to the 121 stations used for the interpolation, and located in various environments, making the test more stringent. Fig. 12 compares the observed and estimated values of the Markov chain parameters and of the gamma parameters for the 7 independent stations. Although large errors occur at a few stations, on the whole the estimates are fair. The correlations with obser-

Fig. 11. Final interpolation of the gamma scale parameter for the month of April (in tenths of mm). Thin white lines: elevation contours (500, 1000 and 2000 m). White: areas where the interpolation cannot be safely performed with the available data

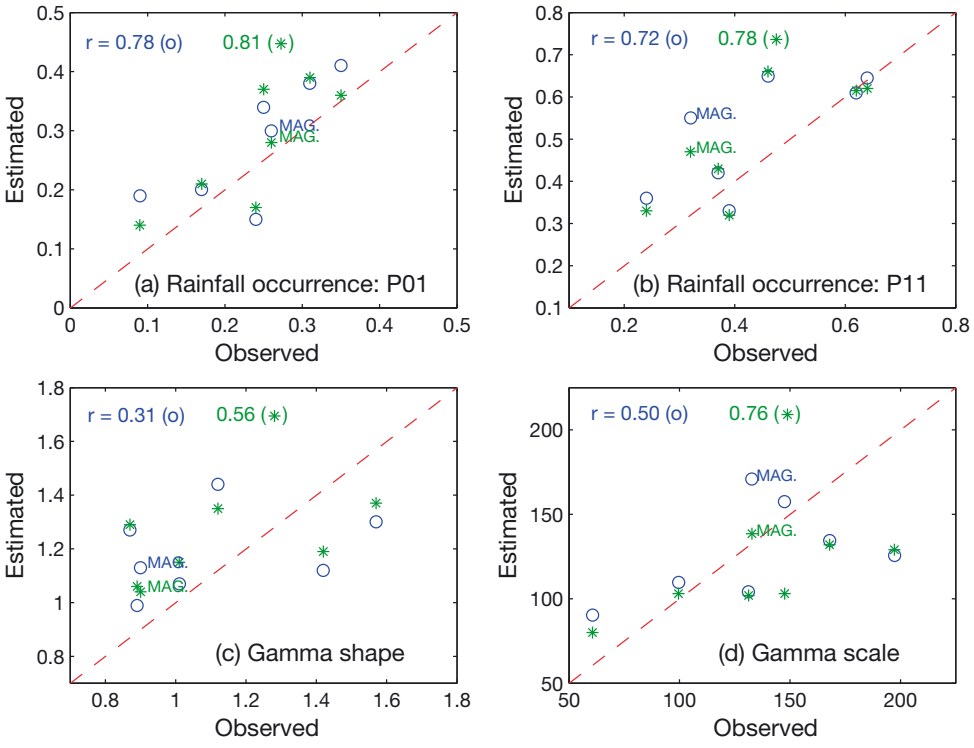


Fig. 12. Observed and estimated values of the Markov-1 and gamma parameters, at 7 independent stations. Blue circles: estimates obtained using the best topographical factor; green stars: using all topographical factors. Values at the top of each panel: correlation coefficients between observed and estimated values for the 2 interpolation methods. MAG: Magadi Stn. Gamma scale values: in tenths of mm

vation are high for the Markov parameters, and lower for the gamma parameters. The case of Magadi, in the dry inner basin of the southern Rift Valley, is enlightening. Although there is no station at all in this peculiar area in the initial data set, the map actually replicates the high gamma scale values (characteristic of infrequent, but heavy daily rainfall events) found at Magadi, although the magnitude is too high. Note that the inclusion of all the topographical factors in the weighted regressions significantly improves the results. The gamma scale parameter at Magadi is reduced to 13.8 mm, down from 17.1 mm when only one topographical factor is used in the model, as compared to 13.3 mm in the observation (Fig. 12).

Statistics on the daily rainfall distribution and the frequency number of wet and dry spells of various durations are also computed for both the observed and reconstructed data. The estimated probabilities of wet and dry spells were obtained from the interpolated values of p01 and p11, following a geometric distribution (Wilks & Wilby 1999). The estimated distribution of daily rainfall amounts was directly obtained from the gamma function with its 2 parameters being derived from the interpolated fields. For Magadi Stn and the month of April (the wettest of the year), Fig. 13a,b compare the observed and estimated (from the reconstructed data) distributions of daily rainfall amounts, and frequencies of dry and wet spells of various durations. Panels a and b are similar except that only one topographical factor has been used in the estimation of the SRG parameters in panel a, while all topographical factors are considered in panel b. Magadi is a relatively dry station, and the reconstructed series underestimate the number of isolated (1 d) wet events (Fig. 13a). Oteng'i & Ogallo (1988) underlined the dominance of 1 d wet spells in dry areas in Kenya. The dry spells are better simulated, although long dry spells tend to be slightly underestimated. The frequency of small rainfall amounts is also strongly underestimated, and that of large amounts overestimated (the 95th percentile stands at 55 mm, against 40 mm in the observation). The model which takes into account all topographical factors (Fig. 13b) considerably improves the results. Although the number of isolated wet days is still underestimated, the daily rainfall amounts are better replicated (including the 95th percentile), the frequency of long dry spells is increased, and the distribution of wet spells is somewhat closer to the observation. For the wetter station of Naivasha (Fig. 13c), the frequency of dry spells is again well replicated. A bias is still found in the frequency of wet spells (too few short spells), but the distribution of rainfall amounts is fairly well reproduced, except in the range of daily events between 15 and 30 mm, which results in an overestimation of the 95th percentile. However, the marked difference in the shape of the observed daily rainfall distribution between

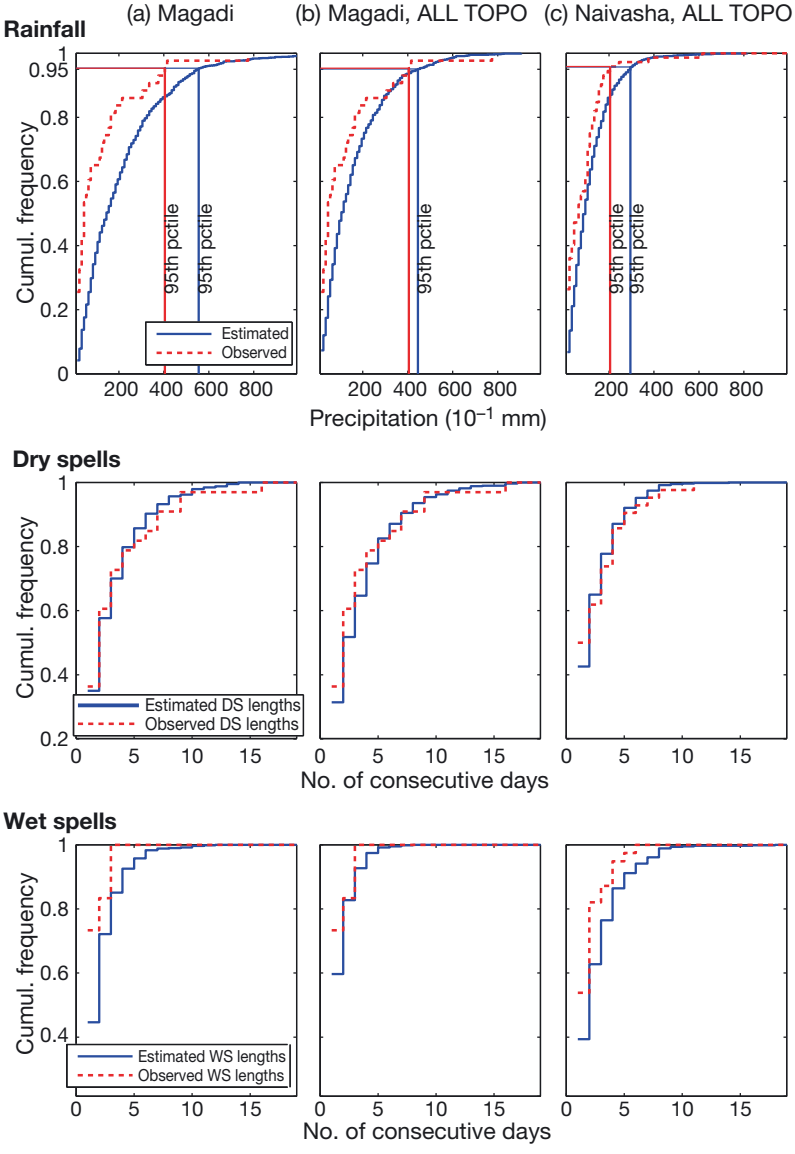


Fig. 13. Observed and reconstructed distribution of daily rainfall amounts and frequencies of wet and dry spells for the month of April at (a,b) Magadi and (c) Naivasha. (a) Only one topographical factor is used in the interpolation; (b,c) all topographical factors (ALL TOPO) are used. Vertical lines: 95th percentile

Magadi (few medium-size events, several heavy events) and Naivasha (many medium-size events, few heavy events) is well reproduced.

5. CONCLUSIONS AND RECOMMENDATIONS

The spatial interpolation of the parameters used in a stochastic rainfall generator, for a topographically complex region of East Africa, was tested using geographically weighted linear regression. The methodology involved the definition of objective geographical variables, including synthetic topographical descriptors. These variables were used for weighing the values of the parameters computed from observed rainfall data at each available station. Local linear regression models were then defined at each station, with the altitude of stations as a predictor, and the geographical variables as well as the inter-station distance as weights. The results were cross-validated and compared to a simpler interpolation scheme, and the resulting maps and models were analysed in order to document the geographical factors explaining the characteristics of daily rainfall in the region. The outcomes of the study are therefore 2-fold: methodological and climatological.

On the methodological side, weighted linear regression is confirmed as a valuable tool for the spatial interpolation of rainfall variables, even in a region displaying a rough topography, and with a data set being far from perfect (both in terms of spatial sampling and data quality). Local regressions, taking into account the nearby environment of each station, outperform models based an equal weighting of all the stations of the domain (domain-wide regression), especially for the Markov chain parameters. This is in line with previous interpolation studies of temperature and precipitation mean fields (Brundson et al. 2001, Joly et al. 2011) or SG parameters (Wilks 2008) for non-tropical regions. Cross-validated scores show that the inclusion of topographical variables as weights further improves the interpolation skill as compared to a standard distance weighting. Both the skills and predictor variables are season-dependent. The main difference is, as expected, between the rainy seasons and the dry seasons. However, there are within season differences (especially for the parameters related to rainfall occurrence) which justify the need to compute the parameters and set the interpolation scheme at a monthly time-scale.

On the climatological side, an important result is the fact that the spatial distribution of the stochastic weather generator parameters is far from random. Part of the spatial patterns of both rainfall occurrence

(as described by Markov chain parameters) and rainfall intensity (as described by the parameters of gamma distribution) is strongly controlled by topography. This includes altitudinal gradients, distance to the sea and terrain shape, especially wind- and leeward effects. In the present study, altitudinal gradients were considered to vary at relatively local scale depending on the other topographical effects. These effects have a noticeable impact on the rainfall intensity, with windward slopes (facing the Indian Ocean) showing greater values of the gamma scale parameter, hence larger amounts of rain per rainday, during most of the year. Rainfall occurrence is also dependant on topography, but with a stronger influence of altitude. Aspect has a very noticeable effect as well. For the Markov-chain p01 parameter for instance, the topographical weighting variable resulting in the smallest interpolation error includes the west/east aspect from May to November, and the south/north aspect from December to April.

On the whole, given the relatively small number of stations available, in a region having such a complex topography, the interpolation scheme does a reasonable job in estimating the parameters of the SG. The use of independent stations located in contrasted environments, even with short records, further demonstrates the skill of the method. In some areas, however, the estimated values of the parameters still display significant residuals. This partly results from the imperfect sampling of daily rainfall data, which leads to uncertainty in the empirical fitting of the Markov chains and gamma parameters at some locations, especially those whose rainfall records only consist of a small number of years. For South Dakota in the US, Woolhiser & Roldan (1986) also found that methodological differences affecting small precipitation amounts, partly related to observation time, accounted for the spatial variability of the parameters from a stochastic model of daily precipitation.

There is still scope for further improvement of the interpolation scheme. For instance, environmental variables other than elevation could be used as local predictors in the weighted regression models. Similarly, the weighting procedure could take into account the quality of the data (e.g. synoptic or non-synoptic stations, long or short times-series). It was found that for the scale parameter of the gamma distribution, which describes the intensity of the rainfall events, there is a weakly significant difference between synoptic and non-synoptic stations, the latter tending to have a small bias towards high intensities, perhaps as a result of the omission of some low rainfall days (not shown).

Another improvement would be to consider interannual variability. As demonstrated for other regions (Jones & Thornton 1997, Wilks 1999, Grondona et al. 2000, Mavromatis & Hansen 2001, Wang et al. 2006), there may be shifts from year to year in the distribution of daily rainfall occurrence and amounts, which imply changes in the parameters of the SGs. Wilks (2002) proposed a method to adjust the parameters to seasonal forecasts given in tercile form. However, Hansen & Ines (2005) showed that without explicitly adjusting the parameters to interannual variations, a good (or even better) reproduction of crop yields can be obtained by simply constraining the disaggregated daily rainfall time-series to match the monthly rainfall total. Given the large interannual variability of rainfall in East Africa, these issues are worth being considered when using the above interpolation scheme for downscaling seasonal forecasts.

Acknowledgements. This study is a contribution to the PICREVAT project, funded by the French National Research Agency (ANR 08-VULN-01-008). Calculations were performed using HPC resources from DSI-CCUB (Université de Bourgogne). Useful comments from the anonymous reviewers also contributed to significantly improve the manuscript.

LITERATURE CITED

Arnold CD, Elliot WJ (1996) CLIGEN weather generator predictions of seasonal wet and dry spells in Uganda. Trans ASAE 39:969–972

Bardossy A, Plate EJ (1992) Space-time model for daily rainfall using atmospheric circulation patterns. Water Resour Res 28:1247–1259

Baron C, Sultan B, Balme M, Sarr B and others (2005) From GCM grid cell to agricultural plot. Scale issues affecting modelling of climate impact. Philos Trans R Soc Lond B Biol Sci 360:2095–2108

Barron J, Rockstrom J, Gichuki F, Hatibu N (2003) Dry spell analysis and maize yields for two semi-arid locations in east Africa. Agric Forest Meteorol 117(1–2):23–37

Becker JJ, Sandwell DT, Smith WHF, Braud J and others (2009a) Global bathymetry and elevation data at 30 arc seconds resolution SRTM30_PLUS. Mar Geod 32: 355–371

Becker EJ, Berbery EH, Higgins RW (2009b) Understanding the characteristics of daily precipitation over the United States using the North American Regional Reanalysis. J Clim 22:6268–6286

Biamah EK, Sterk G, Sharma TC (2005) Analysis of agricultural drought in Iiuni, Eastern Kenya. Application of a Markov model. Hydrol Processes 19:1307–1322

Boulanger JP, Martinez F, Penalba O, Carlos Segura E (2007) Neural network based daily precipitation generator (NNGEN-P). Clim Dyn 28:307–324

Brunsdon C, McClatchey J, Unwin DJ (2001) Spatial variations in the average rainfall–altitude relationship in Great Britain. An approach using geographically weighted regression. Int J Climatol 21:455–466

Camberlin P, Moron V, Okoola R, Philippon N, Gitau W (2009) Components of rainy seasons variability in Equatorial East Africa: onset, cessation, rainfall frequency and intensity. Theor Appl Climatol 98(3–4):237–249

Camberlin P, Boyard-Micheau J, Philippon N, Baron C, Leclerc C, Mwongera C (in press) Climatic gradients along the windward slopes of Mount Kenya and their implication for crop risks. 1. Climate variability. Int J Climatol, doi:10.1002/joc.3427

Castellvi F, Mormeneo I, Perez PJ (2004) Generation of daily amounts of precipitation from standard climatic data. A case study for Argentina. J Hydrol (Amst) 289:286–302

Chapman T (1998) Stochastic modelling of daily rainfall: the impact of adjoining wet days on the distribution of rainfall amounts. Environ Model Softw 13:317–324

Daly C, Neilson RP, Phillips DL (1994) A statistical-topographic model for mapping climatological precipitation over mountainous terrain. J Appl Meteorol 33:140–158

Daly C, Halbleib M, Smith JI, Gibson WP and others (2008) Physiographically-sensitive mapping of temperature and precipitation across the conterminous United States. Int J Climatol 28:2031–2064

Deni SM, Jemain AA, Ibrahim K (2009) Fitting optimum order of Markov chain models for daily rainfall occurrences in Peninsular Malaysia. Theor Appl Climatol 97: 109–121

Fotheringham AS, Brunsdon C, Charlton M (2002) Geographically weighted regression: the analysis of spatially varying relationships. Wiley, Chichester

Geng S, Penning De Vries FWT, Supit I (1986) A simple method for generating daily rainfall data. Agric For Meteorol 36:363–376

Gitau W, Ogallo L, Mutemi JN (2008) Intraseasonal characteristics of wet and dry spells over Kenya. J Kenya Meteorol Soc 2:18–28

Gitau W, Ogallo L, Camberlin P, Okoola R (2013) Spatial coherence and potential predictability assessment of intraseasonal statistics of wet and dry spells over Equatorial Eastern Africa. Int J Climatol 33:2690–2705

Grondona MO, Guillermo PP, Mario B, Monica M, Hugo H (2000) A stochastic precipitation generator conditioned on ENSO phase. A case study in southeastern South America. J Clim 13:2973–2986

Hansen JW, Ines AVM (2005) Stochastic disaggregation of monthly rainfall data for crop simulation studies. Agric For Meteorol 131:233–246

Hession SL, Moore N (2011) A spatial regression analysis of the influence of topography on monthly rainfall in East Africa. Int J Climatol 31:1440–1456

Hills RC (1974) East African rainfall: central tendencies and skewness as synoptic indicators. Weather 9:345–355

Husak GJ, Michaelsen J, Funk C (2007) Use of the gamma distribution to represent monthly rainfall in Africa for drought monitoring applications. Int J Climatol 27: 935–944

Hutchinson P (1990) Frequency distributions of daily, monthly, seasonal and annual rainfalls in Somalia, and their use in the generation of rainfall distributions in data-deficient areas. Third WMO Symp Meteor Aspects Tropical Droughts, Niamey, 30 Apr– 4 May 1990. WMO Tropical Meteorology Research Program Series 36: 239–245

Ines AVM, Hansen JW (2006) Bias correction of daily GCM rainfall for crop simulation studies. Agric For Meteorol 138:44–53

Jimoh OD, Webster P (1996) Optimum order of Markov chain for daily rainfall in Nigeria. J Hydrol (Amst) 185: 45–69

Johnson GL, Daly C, Taylor GH, Hanson CL (2000) Spatial variability and interpolation of stochastic weather simulation model parameters. J Appl Meteorol 39:778–796

Joly D, Brossard T, Cardot H, Cavailhes J, Hilal M, Wavresky J (2011) Temperature interpolation based on local information. The example of France. Int J Climatol 31:2141–2153

Jones PG, Thornton PK (1993) A rainfall generator for agricultural applications in the tropics. Agric For Meteorol 63:1–19

Jones PG, Thornton PK (1997) Spatial and temporal variability of rainfall related to a third-order Markov model. Agric For Meteorol 86:127–138

Jones PG, Thornton PK (1999) Fitting a third-order Markov rainfall model to interpolated climate surfaces. Agric For Meteorol 97:213–231

Jones PG, Thornton PK (2000) MarkSim software to generate daily weather data for Latin America and Africa. Agron J 92:445–453

Kittel TGF, Rosenbloom NA, Royle JA, Daly C and others (2004) VEMAP Phase 2 bioclimatic database. I. Gridded historical (20th) century climate for modeling ecosystem dynamics across the conterminous USA. Clim Res 27: 151–170

Lana X, Burgueño A (1998) Daily dry–wet behaviour in Catalonia (NE Spain) from the viewpoint of Markov chains. Int J Climatol 18:793–815

Larsen GA, Pense RB (1982) Stochastic simulation of daily climatic data for agronomic models. Agron J 74:510–514

Mavromatis T, Hansen JW (2001) Interannual variability characteristics and simulated crop response of four stochastic weather generators. Agric For Meteorol 109: 283–296

Moore TR (1979) Rainfall erosivity in East Africa. Geogr Ann 61:147–156

Nieuwolt S (1974) Rainstorm distributions in Tanzania. Geogr Ann 56A:241–250

Ochola OW, Kerkides P (2003) A Markov chain simulation model for predicting critical wet and dry spells in Kenya. Analysing rainfall events in the Kano Plains. Irrig Drain 52:327–342

Oettli P, Camberlin P (2005) Influence of topography on monthly rainfall distribution over East Africa. Clim Res 28:199–212

Ogallo LA (1993) Dynamics of the East-African climate. Proc Ind Acad Sci Earth Planet Sci 102:203–217

Oteng'i SBB, Ogallo LJ (1988) Persistence of daily rainfall over some parts of Kenya. In WAMEX related research and tropical meteorology in Africa, WMO Tropical Meteorology Research Program Series 28:172–177

Richardson CW, Wright DA (1984) WGEN: a model for generating daily weather variables. USDA/ARS (1984) ARS-8

Robertson AW, Ines A, Hansen JW (2007) Downscaling of seasonal precipitation for crop simulation. J Appl Meteorol Climatol 46:677–693

Said M, Okwi P, Ndenge'e G, Agatsiva J, Kilele X (2007) Nature benefits in Kenya: an atlas of ecosystem and human well-being. World Resource Institute, Department of Resource Surveys and Remote Sensing, Ministry of Environment and Natural Resource, Kenya Central Bureau of Statistics, Ministry of Planning and National

Development, Kenya and International Livestock Research Institute

Semenov MA (2007) Development of high-resolution UKCIP02-based climate change scenarios in the UK. Agric For Meteorol 144:127–138

Semenov MA, Brooks RJ (1999) Spatial interpolation of the LARS-WG stochastic weather generator in Great Britain. Clim Res 11:137–148

Sharma TC (1996) Simulation of the Kenyan longest dry and wet spells and the largest rain-sums using a Markov model. J Hydrol 178(1–4):55–67

Smith RE, Schreiber HA (1974) Point processes of seasonal thunderstorm rainfall. 2. Rainfall depth probabilities. Water Resour Res 10:418–423

Song Y, Semazzi FHM, Xie L, Ogallo LJ (2004) A coupled regional climate model for the Lake Victoria Basin of East Africa. Int J Climatol 24:57–75

Srikanthan R, McMahon TA (2001) Stochastic generation of annual, monthly and daily climate data: a review. Hydrol Earth Syst Sci 5:653–670

Stern RD, Coe R (1984) A model fitting analysis of daily rainfall data. J R Stat Soc [Ser A] A147:1–34

Wang J, Bruce TA, Guido DS (2006) Stochastic modeling of daily summertime rainfall over the southwestern United States. I. Interannual variability. J Hydrometeorol 7:739–754

Wilby RL, Wigley TML, Conway D, Jones PD, Hewitson BC, Main J, Wilks DS (1998) Statistical downscaling of general circulation model output. A comparison of methods. Water Resour Res 34:2995–3008

Wilks DS (1999) Interannual variability and extreme-value characteristics of several stochastic daily precipitation models. Agric For Meteorol 93:153–169

Wilks DS (2002) Realizations of daily weather in forecast seasonal climate. J Hydrometeorol 3:195–207

Wilks DS (2008) High-resolution spatial interpolation of weather generator parameters using local weighted regressions. Agric For Meteorol 148:111–120

Wilks DS, Wilby RL (1999) The weather generation game: a review of stochastic weather models. Prog Phys Geogr 23:329–357

Woolhiser DA, Roldan J (1986) Seasonal and regional variability of parameters for stochastic daily precipitation models. Water Resour Res 22(6):965–978

Appendix. Mathematical description of the methodology

Stochastic rainfall generator (SRG)

The SRG used in the study has 2 parts. The first part simulates rainfall occurrence through a first-order Markov-chain model. A binary daily precipitation variable z is defined, where z(t) = 1 for a wet day and z(t) = 0 for a dry day. The Markov model uses 2 transition probabilities, computed for each station and each month separately:

$$p01 = Pr\{z(t) = 1 \mid z(t-1) = 0\} \quad (A1)$$

and

$$p11 = Pr\{z(t) = 1 \mid z(t-1) = 1\} \quad (A2)$$

The second part of the stochastic model estimates rainfall amount for each wet day using a 2-parameter gamma distribution, whose density function is as follows:

$$g(x,\alpha,\beta) = \frac{x^{\alpha-1} \cdot e^{-x/\beta}}{\beta^{\alpha} \cdot \Gamma(\alpha)} \quad (A3)$$

where x is the daily precipitation amount, Γ is the gamma function, parameter α is the shape parameter (dimensionless) and parameter β is the scale parameter (units of precipitation). These 2 parameters are estimated for each station and each month separately, based on the method of maximum likelihood.

Kolmogorov–Smirnov (KS) test statistics are used to assess the goodness-of-fit of both the Markov chain models (Lana & Burgueño 1998) and the gamma distribution (Larsen & Pense 1982). The KS statistics (at 99% confidence levels) are first used to compare the cumulative distribution functions of the lengths of wet and dry spells (observed and simulated by the Markov chain models), and then to compare the cumulative distribution functions of daily rainfall amounts (observed and simulated from the gamma distribution).

Spatial interpolation of the SRG parameters

The interpolation of each parameter Y of the SRG (p01, p11, α and β, successively) is based on weighted local regressions, using elevation (E) as a single predictor:

$$\hat{Y} = aE + b \quad (A4)$$

The coefficients a and b are allowed to vary in space. At a given pixel, they are obtained by minimizing the weighted sum of squared residuals, WSS:

$$WSS = \sum_{i=1}^{s} W_i \cdot (Y_i - \hat{Y}_i)^2 \quad (A5)$$

with s the number of nearby stations used in the regression, Y_i the observed value of the parameter at station i and W_i the weight attached to each station i.

The weights (W) are computed based on a variety of geographical properties, considered separately or combined. The first set of weights $WDIST_i$ considers the horizontal distance D_i between each station i and the target pixel, as follows:

$$WDIST_i = \frac{K_i}{\sum_{i=1}^{s} K_i} \quad (A6)$$

where

$$K_i = \frac{15}{16}\left(1 - \frac{D_i^2}{D_{MAX}^2}\right)^2 \quad (A7)$$

is a biweight kernel function, as in Wilks (2008). D_{MAX} is a predefined distance threshold considered as the radius of influence around each location. Eqs. (A6) & (A7) mean that a WDIST weight of zero is assigned to stations located at a distance to the pixel exceeding the predefined threshold

Appendix (continued)

D_{MAX}. The weights gradually increase from zero to one as the station comes closer to the target pixel. The optimal value of D_{MAX} is chosen in the range of 50 to 200 km, and set through cross-validation.

In case the number of stations available within the initial D_{MAX} radius is below a predefined number (NS), the value of D_{MAX} is gradually increased, in steps of 0.5°, until the required number of stations is reached. Several tests have been done to determine the value of NS. Due to a low density of stations in some parts of the region, and in order to conveniently reflect small-scale climate patterns in these areas, NS was finally set to a low value (4 stations), but in most cases the number of available stations used in the regressions is >10.

The second set of weights considers the distance between the pixel and each station in terms of coastal proximity (COAST; Table 1). Two groups of stations are defined: subcoastal stations, located at a maximum distance DC_{MAX} to the sea, and inland stations. If the target pixel is a subcoastal one, a maximum weight of 1 is given to the stations which are located at the same distance to the sea as the pixel. Weights then linearly decrease for the other stations, as a function of the difference in coastal nearness between the target pixel and each station:

$$W\mathrm{COAST}_i = \begin{cases} 0 & if\ \ \mathrm{DSEA}_i > \mathrm{DC}_{\mathrm{MAX}} \\ 1 - |(\mathrm{DSEA}_i - \mathrm{DSEA}_j)| / \mathrm{DC}_{\mathrm{MAX}} & if\ \ \mathrm{DSEA}_i < \mathrm{DCMAX} \end{cases} \quad (A8)$$

where DSEA_i is the distance from station i to the sea, and DSEA_j the distance from pixel j to the sea.

For inland pixels (i.e. at a distance to the sea $>\mathrm{DC}_{\mathrm{MAX}}$), the influence of the sea is considered as negligible, and all the stations apart from the subcoastal ones are given the same weight of 1. The subcoastal stations are then given a weight between 0 and 1 and linearly related to the distance to the sea, as follows:

$$W\mathrm{COAST}_i = \begin{cases} 1 & if\ \ \mathrm{DSEA}_i > \mathrm{DC}_{\mathrm{MAX}} \\ \mathrm{DSEA}_i / \mathrm{DC}_{\mathrm{MAX}} & if\ \ \mathrm{DSEA}_i < \mathrm{DC}_{\mathrm{MAX}} \end{cases} \quad (A9)$$

As for D_{MAX}, the optimal value of DC_{MAX} is chosen in the range of 50 to 300 km through cross-validation.

The last set of weights describes the topographical setting of the pixel and how similar it is to that of the stations used in the linear regression. To describe the topographical setting, 8 variables (TOPO_1 to TOPO_8) are used which account for slope aspect, relief geometry, and average slope The initial values of TOPO_{1-7} range between –1 and 1 (see details in Table 1, and Oettli & Camberlin 2005 on how they were obtained). The corresponding weights are computed as follows:

$$W\mathrm{TOPO}_{ijk} = 1 - \frac{\mathrm{DTOPO}_{ijk}}{\max(\mathrm{DTOPO}_{jk})} \quad (A10)$$

where

$$\mathrm{DTOPO}_{ijk} = \sqrt{|(\mathrm{TOPO}_{ik} - \mathrm{TOPO}_{jk})|} \quad (A11)$$

with TOPO_{ik} as the value of the topographical variable k at station i, and TOPO_{jk} as the value of the same topographical variable at pixel j. Eq. (A11) is a measure of similarity of the topography between the pixel and the station, and Eq. (A10) standardises the topographical weights between 0 (fully dissimilar topographical settings) and 1 (perfectly similar topographical settings).

As in Oettli & Camberlin (2005), all these variables are computed using 4 different window sizes around the pixel (9, 39, 123 and 213 km). The optimal window size chosen is set through cross-validation.

Different interpolation models are defined in which the weights differ. A benchmark model is defined where all weights are set to 1, which is equivalent to a domain-wide linear interpolation using altitude as the predictor. Other models are defined where only one set of weights is turned on (e.g. DIST, COAST, etc.). More elaborated models are produced where different sets of weights are combined. In this case, the final weights (W) are obtained as the product of the different sets of weights, as for example in the full model:

$$W = W\mathrm{DIST} \cdot W\mathrm{COAST} \cdot \prod_{k=1}^{8} W\mathrm{TOPO}_k \quad (A12)$$

Leave-one-out cross-validation is carried out in order to assess the general performance of the models and to optimally assign the values of the parameters used in the weighting process (D_{MAX}, DC_{MAX} and size of the windows used to define topographical features).

Editorial responsibility: Bryson Bates, Wembley, Western Australia, Australia

Submitted: February 6, 2013; Accepted: September 23, 2013
Proofs received from author(s): January 22, 2014