

Use of the gamma distribution to represent monthly rainfall in Africa for drought monitoring applications

Gregory J. Husak,* Joel Michaelsen and Chris Funk

Climate Hazard Group, Department of Geography, University of California, Santa Barbara, CA 93106, USA

Abstract:

Evaluating a range of scenarios that accurately reflect precipitation variability is critical for water resource applications. Inputs to these applications can be provided using location- and interval-specific probability distributions. These distributions make it possible to estimate the likelihood of rainfall being within a specified range. In this paper, we demonstrate the feasibility of fitting cell-by-cell probability distributions to grids of monthly interpolated, continent-wide data. Future work will then detail applications of these grids to improved satellite-remote sensing of drought and interpretations of probabilistic climate outlook forum forecasts. The gamma distribution is well suited to these applications because it is fairly familiar to African scientists, and capable of representing a variety of distribution shapes. This study tests the goodness-of-fit using the Kolmogorov–Smirnov (KS) test, and compares these results against another distribution commonly used in rainfall events, the Weibull. The gamma distribution is suitable for roughly 98% of the locations over all months. The techniques and results presented in this study provide a foundation for use of the gamma distribution to generate drivers for various rain-related models. These models are used as decision support tools for the management of water and agricultural resources as well as food reserves by providing decision makers with ways to evaluate the likelihood of various rainfall accumulations and assess different scenarios in Africa. Copyright © 2006 Royal Meteorological Society

KEY WORDS rainfall; gamma distribution; Africa; precipitation; Kolmogorov–Smirnov; drought

Received 21 December 2005; Revised 11 September 2006; Accepted 16 September 2006

INTRODUCTION

In order to improve the ability of African decision makers to prepare for and deal with the consequences of precipitation anomalies, it is important to provide them with a more complete understanding of the range and likelihood of rainfall totals a location could possibly receive. Models of rainfall probability distributions over various timescales are useful tools for gaining this kind of understanding. Modeling rainfall variability in Africa presents an imposing problem for many reasons, including the need to summarize rainfall data for many years at many sites and the difficulty in finding a single method to represent such a variety of rainfall regimes. Rainfall regimes across the vast African continent differ widely in terms of total accumulations, seasonal timing, and amounts of variability. Any method that is applicable across this wide range of conditions has to be quite flexible.

The typical approach to gaining a better understanding of the spatial and temporal variability in precipitation starts with the acquisition of historical rainfall data. These historical data provide necessary information about accumulation amounts in both time and space for the region

and form the basis for fitting and testing distribution models. When historical data is unavailable in a region, or available data is inaccurate or incomplete in a spatial or temporal sense, geophysical models can be used to ‘fill in’ the missing values. These geophysical models are based on available data at other locations and times, as well as additional variables that add information to the model. Assuming the historical values – recorded or modeled – exist and are accepted as reasonably accurate, it is possible to fit parametric statistical distribution models to rainfall histories at individual locations of interest. Using a parametric distribution model allows for a more stable and extensive analysis of the rainfall probabilities than would be available using the raw data directly. The resulting distributions describe the estimated probability of different amounts of rain at a location for a select time interval (e.g. annual, seasonal or monthly), based on the historical values for that interval at that location.

There are many probability distributions that could be successfully utilized to parameterize rainfall distributions. The critical component for these distributions is that they be flexible enough to represent a variety of rainfall regimes. From a practical point of view, there is little difference between many of the commonly used distributions when estimating parameters based on a limited number of points, as is the case in much of the developing world. Of these available distributions, the gamma

* Correspondence to: Gregory J. Husak, Climate Hazard Group, Department of Geography, University of California, Santa Barbara, CA 93106, USA. E-mail: husak@geog.ucsb.edu

distribution is one of the more widely understood, making it a good choice for implementation of this work in developing countries.

Once the parameters of the distribution have been estimated, they can be used to describe rainfall regimes and be used in a variety of applications. The distribution parameters might complement or even replace such common measures as the median, variance, minimum, maximum and quartile values as descriptors of the rainfall at any location. For locating potential hazard hotspots, distribution parameters may be used to identify areas with a disposition towards certain precipitation related hazards such as drought, flood, outbreak of disease, or reliability in providing adequate water for rain-fed agriculture. Monitoring of rainfall conditions may use distribution parameters as the foundation for the standardized precipitation index. Combining distributions with probabilistic forecasts may result in a quantitative estimation of seasonal rainfall accumulations.

The primary objectives of this research are to estimate and evaluate distribution parameters that may be used to describe the probability of monthly rainfall accumulation for a location. More specifically, probability distribution parameters are estimated from monthly model-derived historical rainfall values with a spatial resolution compatible with current agroclimatic models, and the goodness-of-fit of the parameters assessed. Direct interpretation of these parameter estimates results in a broad technique for the description of general rainfall regimes for the entire continent. The result of this research is a new portrayal of African rainfall utilizing these probability distribution parameters. In addition to the interpretation of these parameters, description of potential uses of the parameters in hydrologic resource modeling is introduced. Probabilistic information can be utilized to dynamically evaluate rainfall accumulations as well as test different scenarios that can be used as input into other models.

This paper quantifies the accuracy of gamma distribution grids actively used in applications for drought monitoring in sub-Saharan Africa, where nearly one-third of the population is undernourished (FAO, 2005a). Notably, this work is currently in use as part of the standardized precipitation index available at the Africa Data Dissemination Service (<http://earlywarning.usgs.gov/adds/>) as well as with probabilistic forecasts produced by Climate Outlook Forums or by agencies such as the International Research Institute. The gridded gamma distributions described below are used to provide quantitative interpretations of these forecasts, providing actionable information for food security decisions (<http://iri.columbia.edu/africa/project/FSOFsGHA/>).

DATA AND BACKGROUND

The collaborative historical African rainfall model (CHARM)

In this study, the distributions are fit to historical modeled rainfall at individual cells of the Collaborative

Historical African Rainfall Model (CHARM), a gridded dataset developed to compensate for relatively poor spatial coverage of reliable station data in Africa (Funk *et al.*, 2003).

The CHARM represents the synthesis of historical reanalysis fields, a continental-scale digital elevation model (DEM), and precipitation values based on interpolated station data. The reanalysis data provide a historically deep – if spatially coarse – dataset that allows the CHARM to derive daily rainfall values for a recent 36-year period (1961–1996). The reanalysis data is the result of applying a state-of-the-art data assimilation process to a rich library of quality controlled historical data (Kalnay *et al.*, 1996), and is used to disaggregate monthly totals within the days of the month. The DEM provides a higher spatial resolution (0.1 degree) than the reanalysis data, and allows for the introduction of orographic processes in the model (Funk and Michaelsen, 2004). This resolution is also consistent with other satellite-based rainfall fields and related products routinely used in African drought monitoring activities. Finally, the climatologically aided interpolated (CAI) rain gauge monthly accumulations developed by Willmott and Robeson (1995) are used to constrain the CHARM data so that monthly accumulations of the daily CHARM data match the CAI values in areas where there is no orographic enhancement of rainfall. The CAI data are a well-documented, gauge-based gridded dataset accumulated to monthly values. The result of these complementary datasets is a set of daily rainfall grids with 0.1-degree resolution for the years 1961–1996. Since the CHARM is matched to the monthly station-based grids, the most reliable input dataset, it follows that the monthly CHARM accumulations be used for this research, although analysis at shorter timescales is possible.

The resulting CHARM data reflects some of the characteristics of the input data. Variability in rainfall is relatively low in areas of the continent far from stations (Funk and Michaelsen, 2004), for instance. However, even with these shortcomings, the CHARM data is a good fit for this project owing to the spatial, temporal and financial characteristics of the dataset.

Utilization of the gamma distribution function

The representation of the likelihood of receiving a specific rainfall amount based on 36 (1961–1996) observations for a given interval is best accomplished by fitting a parametric statistical distribution (Ison *et al.*, 1971; Woolhiser, 1992). This parameterized distribution is a continuous function allowing for a comprehensive analysis of the rainfall based on the acquired sample.

With the available 36-year modeled history of rainfall data, it is possible to fit parameterized statistical distributions to the data, but first an appropriate distribution must be selected. Much research has been carried out related to fitting and evaluating statistical distributions for rainfall. Juras (1994) discusses a number of studies that employed a variety of statistical distributions

in an attempt to accurately fit precipitation data. These distributions included the compound Poisson-exponential distribution: the log, square root and cube-root normal distributions, and the gamma distribution. In another such summary, Woolhiser (1992) reviews studies applying various normalizing transforms, as well as the kappa and Weibull distributions. Legates (1991) compared eight different statistical distributions for their goodness-of-fit to station data from around the globe. This study found the Box-Cox transformation to be a superior probability distribution function to employ for assessment of monthly rainfall values. There seems to be ambiguity in the available literature regarding the superior distribution to use when attempting to represent monthly rainfall values.

The gamma distribution is frequently used to represent precipitation because it provides a flexible representation of a variety of distribution shapes while utilizing only two parameters, the shape and the scale (Wilks, 1990). Six example gamma probability distribution functions, each with a mean of 20, are plotted in Figure 1 to illustrate the variety of shapes captured by the gamma distribution. In addition to the gamma parameters, this research includes a third parameter to describe the probability of zero rainfall during the interval. A more complete description of the estimation of the gamma distribution parameters is given in *Estimation of distribution parameters*.

The gamma distribution is a good choice for describing precipitation values for a variety of reasons. The first advantage of the gamma distribution is that it is bounded on the left at zero (Thom, 1958; Wilks, 1995). This is important for precipitation applications because negative rainfall is an impossibility, so a distribution that excludes negative values is readily applicable. This is especially important in dry areas or locations with high variability but a low mean. Second, the gamma distribution is positively skewed, meaning that it has an extended tail to the right of the distribution. This is advantageous because

it mimics actual rainfall distributions for many areas where there is a non-zero probability of extremely high rainfall amounts, even though the typical rainfall may not be very large (Ananthakrishnan and Soman, 1989). Finally, the gamma distribution offers a tremendous amount of flexibility in the shape of the distribution function (Wilks, 1995). The gamma distribution may range from exponential-decay forms for shape values near one, to nearly normal forms for shape values beyond 20 (Wilks, 1990; Öztürk, 1981). This flexibility allows for the gamma distribution to be fit to any number of rainfall regimes with reasonable accuracy, while other distributions may fit only a single, specific rainfall regime.

Many studies have employed the gamma distribution in the analysis of rainfall. Ison *et al.*, (1971) examined the relationship between the gamma distribution parameters and rainfall accumulation period for three rain stations in Kansas, USA. This study showed that the gamma distribution parameters can be scaled to describe rainfall for events of different duration. This is valuable because it means that the gamma distribution is useful for studying rainfall at a variety of timescales from multi-day accumulations to seasonal accumulations.

The accuracy of the estimated gamma distribution (with probabilities calculated from the estimated gamma parameters) in matching the empirical distribution (the empirical probability function based on the observations), can be measured by comparing the cumulative distribution functions of the estimated gamma and empirical distributions (Wilks, 1995). In this research, maximum likelihood estimators (MLEs) are employed to calculate the shape and scale parameters for the gamma distribution. An alternative to the MLE parameters is the method of moments estimation. It has been shown, however, that the method of moments is a poor estimator, owing to inefficiency, for small shape values (Wilks, 1990; Wilks,

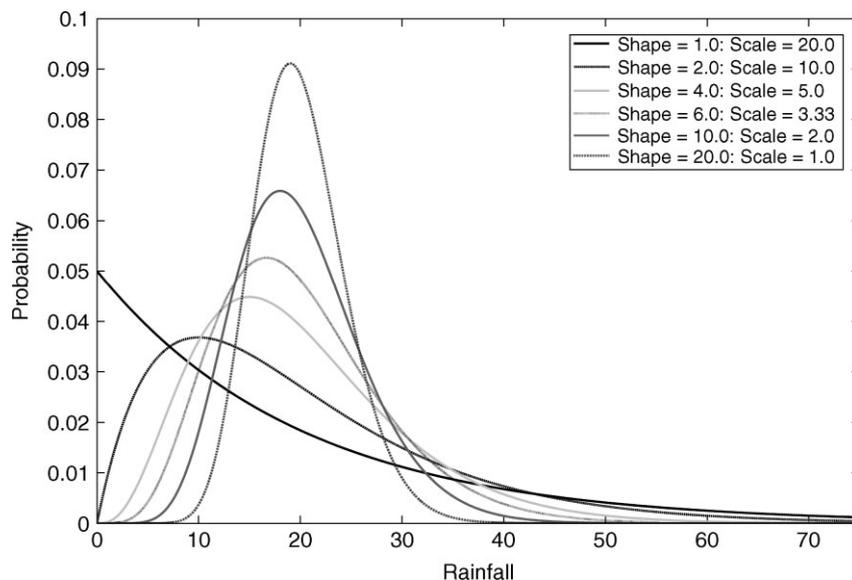


Figure 1. A plot of six unique gamma distribution functions, all having a mean equal to 20. The plot shows the variety of distribution shapes that can be represented by the gamma distribution.

1995; Thom, 1958). Since this research seeks to accurately fit as many regimes as possible, the MLE method is used so as not to eliminate regions with small-scale parameters.

The Kolmogorov–Smirnov test

Once the parameters are estimated, their accuracy in approximating the true rainfall distribution must be confirmed. To do this, the estimated gamma distribution is compared against the empirical distribution (Ison *et al.*, 1971). This can be done with the Kolmogorov–Smirnov (KS) goodness-of-fit test (Crutcher, 1975). Because the values being tested in this scenario are the same values being used in deriving the distribution parameters, this test is sometimes also known as the *Lilliefors test* (Wilks, 1995). The KS test compares the cumulative distribution functions of the theoretical distribution – the distribution described by the estimated shape and scale parameters – with the observed values and returns the maximum difference between these two cumulative distributions (Wilks, 1995). This maximum difference in cumulative distribution functions is frequently referred to as the KS-statistic.

In this statistical test, the null hypothesis is that the observed data are drawn from the chosen theoretical distribution. If the value of the KS-statistic is excessively large, then the null hypothesis is rejected. A rejection would imply that the distribution parameters are not doing an adequate job of modeling the empirical distribution of rainfall at a location. The acceptable KS value for rejection depends on the number of points in the empirical distribution being used to test the theoretical distribution, and the rejection level chosen for the study (Crutcher, 1975). Because the acceptable value of the KS-statistic is variable, the confidence in accepting or rejecting the theoretical distribution may be measured by the *p*-value, which incorporates the number of values being used in the test into the calculation of its value. A small *p*-value is cause for rejection of the null hypothesis, while a *p*-value larger than the selected significance level means that the null hypothesis cannot be rejected (Wilks, 1995).

METHODOLOGY

Incorporating the occurrence of no rainfall

The general gamma distribution does not allow for values less than or equal to zero in the distribution, so the probability of an event with no rainfall must be treated separately. This conditional distribution – accumulation probabilities conditional on the presence of rainfall – is combined with a mixture coefficient used to account for the probability of no rain to create the probability distribution. Values of zero in the rainfall history are initially excluded from the calculation of the shape and scale parameters. In order to account for the occurrence of no rain in the modeled history, this study uses an additional parameter (*q*) in the theoretical distribution corresponding to the probability of receiving no rainfall during the

interval. This probability of no rainfall is estimated by counting the number of occurrences of zero (n_o) and dividing it by the number of historical observations (n), as shown in Equation (1). Since the CHARM data has a 36-year history, n is equal to 36 for each single-month, but may be less if a multi-month interval spans the December to January period. The number of non-zero observations in the n historical records, n_p , used in the estimation of the gamma distribution parameters is defined as $(n - n_o)$ so that only the number of years with positive accumulations are used in the gamma parameter estimation. Additionally, the observations of zero are excluded from the calculation of the mean and sum of natural log terms as described in *Estimation of distribution parameters*.

$$\hat{q} = \frac{n_o}{n} \quad (1)$$

If the number of zero values for a given geographic location and interval is large, there are few non-zero values available to be used in the estimation of the gamma distribution parameters. As \hat{q} increases from 0.0, the reliability of the fit of the theoretical distribution to the empirical values decreases. For example, at geographic locations where \hat{q} is larger than 0.5, the gamma distribution parameters are being estimated by fewer than 18 samples (out of 36 years) and the parameter reliability becomes quite suspect. Research applying estimated distribution parameters suggests that over 30 years of data be used in the estimation of parameters (McKee, 1993; Hayes *et al.*, 1999; Wu *et al.*, 2001).

Estimation of distribution parameters

The gamma probability distribution function is shown in Equation (2), where α is called the *shape parameter* and β is called the *scale parameter*, and x represents a rainfall amount. The complete gamma function ($\Gamma(\alpha)$) (Equation (3)) can be solved or estimated from tables, which can be found in most software packages.

$$f(x) = \frac{(x/\beta)^{\alpha-1} e^{-x/\beta}}{\beta \Gamma(\alpha)} \quad (2)$$

$$\Gamma(\alpha) = \int_0^{\infty} e^{-t} t^{\alpha-1} dt \quad (3)$$

For this study, the gamma distribution parameters are estimated through maximum likelihood estimation. The calculation of the MLEs begins with the calculation of an intermediate value A as shown in Equation (4) (Wilks, 1995; Ozturk, 1981; Thom, 1958), where x_i is equal to all non-zero values in the rainfall history, and the mean (\bar{x}) is the arithmetic mean of all non-zero values. The value A then, is equal to the natural log of the mean minus the mean of the natural logs of the non-zero accumulations at a point. Ultimately, A is a measure of the skewness of the distribution. This value is then used in the estimation of the shape parameter, represented by $\hat{\alpha}$ (Equation (5)). The scale estimator, $\hat{\beta}$, is then the mean divided by the estimated shape parameter (Equation (6)). Finally, the

product of the shape ($\hat{\alpha}$) and the square of the scale ($\hat{\beta}^2$), is approximately equal to the variance (s^2), given as the mean of the sum of squared difference from the mean (Equation (7)).

$$A = \ln(\bar{x}) - \frac{\sum_{i=1}^{n_p} \ln(x_i)}{n_p} \quad (4)$$

$$\hat{\alpha} = \frac{1}{4A} \left(1 + \sqrt{1 + \frac{4A}{3}} \right) \quad (5)$$

$$\hat{\beta} = \frac{\bar{x}}{\hat{\alpha}} \quad (6)$$

$$\hat{\alpha}\hat{\beta}^2 \approx s^2 \quad (7)$$

Analysis of these equations can be helpful in the interpretation and understanding of how the rainfall parameters describe the rainfall. Equation (6) can be rewritten to show that the product of the parameter estimates is equal to the mean of the non-zero values in the rainfall history. The rewriting of Equations (6) and (7) can aid in providing some intuitive understanding of the rainfall distribution at any point.

These calculations are performed for each 0.1-degree by 0.1-degree grid cell in the CHARM data array. The results can be displayed as maps of the estimated parameters for each monthly accumulation period. These MLE parameters describe the rainfall probability distribution of the CHARM dataset at each grid cell for non-zero rainfall amounts.

Implementation of the KS test

Once the gamma distribution parameters are estimated, they should be evaluated to understand how accurately they reflect the historical data, and therefore represent the modeled probability of rainfall for a location. The KS test offers a straightforward method for assessing the relationship between the empirical distribution and the estimated distribution, leading to either acceptance or rejection of the null hypothesis.

As used here, the null hypothesis of the KS test is that the sample – the rainfall history in this case – is taken from the theoretical gamma distribution, with parameters as estimated in Equations (4)–(6). Selection of an acceptable rejection level is arbitrary. For this research, we use a rejection level of 0.10, meaning that we reject the null hypothesis that the theoretical distribution is performing adequately in modeling the historical values, at locations with p -values less than 0.10.

The KS test compares the empirical and theoretical cumulative distributions and returns the absolute value of the largest difference between them. This is mathematically described in Equation (8) (Wilks, 1995). In Equation (8), D_n represents the maximum difference between the empirical and theoretical distributions over all real numbers y , and is referred to as the KS value; $F_n(y)$ is

the empirical cumulative probability of observing a value less than or equal to y as shown in Equation (9), where $1/n_p$ is added for each observation (y_i) that is greater than zero and less than or equal to y ; $F(y)$ is the theoretical cumulative probability at y described by the estimated gamma distribution parameters ($\hat{\alpha}$, $\hat{\beta}$) shown in Equation (10).

$$D_n = \max_y |F_n(y) - F(y)| \quad (8)$$

$$F_n(y) = \frac{\#\{i \in \{1, 2, \dots, n\} : y_i \leq y\}}{n} \quad (9)$$

$$F(y) = \int_0^y f(x)dx = \frac{1}{\hat{\beta}^{\hat{\alpha}}\Gamma(\hat{\alpha})} \int_0^y x^{\hat{\alpha}-1} e^{-x/\hat{\beta}} dx \quad (10)$$

Intuitively, a smaller value of D_n means a better fit between the observed and theoretical distributions for a fixed number of observations (n). Also, having a larger number of points in the empirical distribution usually creates a smoother cumulative function curve and eliminates the large vertical jumps associated with empirical curves based on only a few points. The p -value measures the confidence that the empirical data are taken from the theoretical distribution based on the D_n result and factoring in the number of samples into its calculation. In other words, it is a measure of the probability of achieving a test statistic, D_n , at least as large as the observed measure assuming the null hypothesis, given the number of samples used in estimating the parameters. A location with more samples in the historical dataset has a smaller KS test statistic (D_n) associated with the same p -value. Similarly, if two locations have the same KS test statistic value, the one with fewer samples has a larger p -value (Wilks, 1995; Siegel, 1956).

The KS test was applied at every pixel in the CHARM grid which recorded at least two non-zero rain events during the monthly interval. The exact number of points tested changes depending on the month because the number of gridcells that receive no rainfall in a month varies throughout the year. The estimated parameters at each site were used to estimate the theoretical distribution and compared to the empirical cumulative distribution function from the CHARM monthly rainfall from the same location.

RESULTS AND DISCUSSION

This section highlights the important findings of this research and describes the relevant implications for understanding rainfall in Africa. Since this study relies heavily on modeled data for the results, a note about this would be appropriate. Any use of the CHARM dataset must acknowledge the limitations of the model, which should be used to guide the application of the dataset. The CHARM model is designed to estimate precipitation by combining various inputs to create an output

representing a synergism of the positive characteristics of the input datasets. It is not designed to be spatially explicit, and therefore quantitative analysis should be performed for areas rather than at individual points. The CHARM has been shown to approximate precipitation with reasonable accuracy (Funk *et al.*, 2003). With these limitations in mind, caution must be used in the interpretation of the outputs presented here. Another related issue is the systematic impact that station distribution has on the temporal variance of interpolated fields. Most interpolation schemes employ, in essence, some type of weighted average of the points surrounding a location. This averaging influences the temporal variance in complex ways, with locations near stations typically varying more from time-step to time-step. Still, gridded data and gridded distributions are useful in many applications.

Interpretation of the distribution parameters

Proper interpretation of the gamma distribution parameters requires some understanding of the distribution properties. Unlike the normal distribution where a single parameter, such as the mean or standard deviation can directly provide an intuitive understanding of some aspect of the distribution, the gamma distribution requires that both the shape and scale parameters be interpreted together. Areas with similar shape values, but different scale values, have very different probability density functions describing the rainfall.

A conceptual understanding. The shape parameter describes the form of the curve. Distributions with a low shape parameter are positively skewed, and as the shape value increases the distribution curve becomes more symmetrical. Rewriting Equations (6) and (7) of the gamma distribution, it is possible to find that the square root of the estimated shape parameter is equal to the mean of the non-zero observations divided by the standard deviation (i.e. the inverse of the coefficient of variation). In general terms, this means that a wet area (i.e. one

with a large mean) with a relatively small variance has a large estimated shape parameter, while a dry area with a relatively high variance results in a small estimated shape parameter.

In estimating the scale parameter earlier, we used Equation (6), which can be rewritten to show that the product of the estimates of the shape and scale is equal to the mean of non-zero observations (\bar{x}), as shown in Equation (11). The variance of the gamma distribution is the estimated shape ($\hat{\alpha}$) multiplied by the square of the estimated scale ($\hat{\beta}$) parameter (Equation (7)). Manipulation of these two equations reveals that the scale ($\hat{\beta}$) parameter is approximately the variance divided by the mean (Equation (12)). So large-scale ($\hat{\beta}$) values may indicate relatively high variance in a dry region (low mean), and small-scale ($\hat{\beta}$) values may indicate relatively little variance in wet areas.

$$\hat{\alpha}\hat{\beta} = \bar{x} \quad (11)$$

$$\hat{\beta} \approx \frac{s^2}{\bar{x}} \quad (12)$$

The mapped parameters provide some spatial context to the rainfall values and distributions. In the evaluation of the parameter fields, it becomes apparent that areas receiving a minimal amount of rainfall are described by either a large shape parameter or a large-scale parameter, but not large values in both parameters. This discussion uses the term ‘shape-dominated’ rainfall to refer to locations with a larger shape parameter, and ‘scale-dominated’ rainfall to refer to locations with a larger scale parameter. Figure 2 shows a conceptual graphic with the shape and scale parameters on the x-axis and y-axis, respectively. The axes on this graphic are not numerated as the concept of ‘small’ and ‘large’ values may vary in time and space. This graphic provides an idea of how the parameters describe regimes. Areas in the ‘low rainfall’ region of the plot describe regions that are typically arid during the interval of analysis. The empty area in the

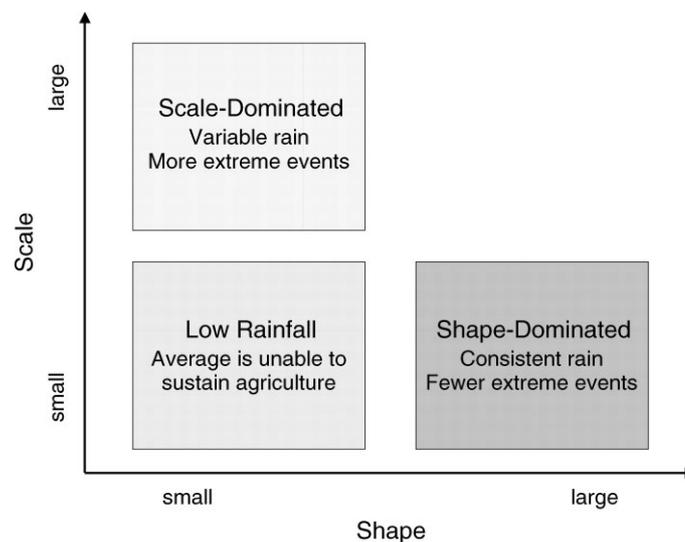


Figure 2. Conceptual regimes described in parameter space with the shape parameter (α) on the x-axis and the scale parameter (β) on the y-axis.

graph indicates that areas with at least a minimal amount of rainfall are in either the ‘shape-dominated’ or ‘scale-dominated’ category.

A shape-dominated regime describes a pattern where the rainfall tends to be symmetrically distributed, indicating that drier-than-average events are as common as wetter-than-average events. This tends to describe areas that typically receive consistent rainfall accumulations in the historical record. Additionally, using Equation (6), if mean rainfall is held constant, a larger $\hat{\alpha}$ value results in a smaller $\hat{\beta}$, meaning less variance in the distribution function according to Equation (7). For areas with similar means, a shape-dominated regime indicates a tighter distribution of rainfall around the median than a scale-dominated regime.

Scale-dominated rainfall describes locations where the variance is quite large in comparison to the mean. Again, if the mean rainfall is held constant, as the scale increases the shape parameter must decrease resulting in a more positively skewed distribution function (from Equation (6)). However, if the shape parameter is held constant, as the scale increases there is a larger mean along with a larger variance. This illustrates that both the shape and scale parameters must be interpreted when comparing the rainfall distributions of two locations, as the comparison of only a single parameter can lead to erroneous conclusions about the rainfall at each place.

A spatial analysis. Interpretation of the parameter grids, as shown for January in Figure 3, yields some general observations regarding the spatial distribution of the parameters.

Large shape values tend to follow the Inter-tropical Convergence Zone (ITCZ) through each of the monthly maps. This makes intuitive sense because these locations receive relatively large and consistent rainfall when the ITCZ is present. In these instances, the increase in the average rainfall would lead to a lower variance relative to the mean (but not necessarily a lower absolute variance), and since the product of the shape and scale must equal the mean, if the scale decreases then the shape must increase, in order to preserve a constant mean.

A high shape value also means that the distribution curve is getting more symmetrical, as shown in Figure 1, and that the probability of events drier-than-average is approximately equal to the probability of events wetter than average.

In addition to large shape values in the ITCZ, there also appear to be some artificially inflated values in areas surrounding major desert regions. These locations are areas with a high probability of no rain, and the parameters are calculated based on only a few positive values. If the small number of years receiving rain received the same or nearly the same accumulation during those wet years then the alpha parameter is artificially inflated as the value A from Equation (4) approaches zero and results in a large shape value. These locations are represented by a low-variance, bell-shaped distribution where all probabilistic realizations of rain, based on the historical data, are covered by only a few values. It is possible to screen these locations to remove any potential for misunderstanding, but by using the spatial context as well as the information provided by the other parameters it is possible to understand what occurs at these locations without misleading the user.

Scale-dominated rainfall frequently appears on the perimeter of the ITCZ where rainfall is more variable. This variability is largely owing to fluctuations in the procession of the ITCZ across the continent. As the ITCZ moves throughout the season, it may stall in some locations for extended periods of time, and then advance quickly through other regions. These irregular movements may give one location heavy rainfall for an extended period of time while causing another to have an abbreviated rainy season. The affect of a location receiving some wet years and some dry years leads to a large variance in the sample, and thus a large-scale value.

Implications for the occurrence of rainfall. By combining the conceptual figure (Figure 2) with the mapped parameters (Figure 3), it is possible to create a classified map based on simple rules that can locate areas where occasional drought may have an impact on agriculture in

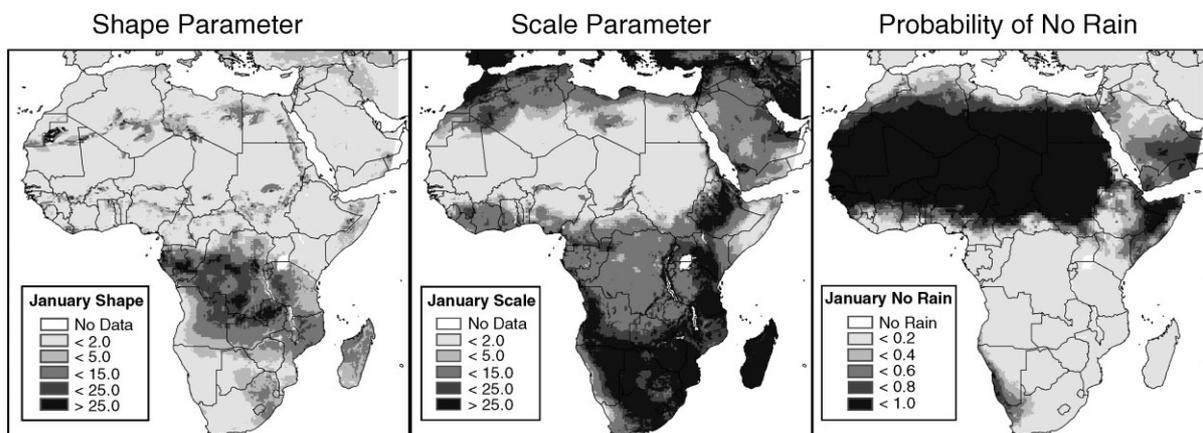


Figure 3. The shape (left), scale (center) and no rain (right) parameters for the month of January.

southern Africa. This exercise is designed as a straightforward example of an application of distribution parameters in defining potential hazard zones. Four classes were created defining extremely wet areas (mean January rainfall over 250 mm), regularly dry areas (mean January rainfall less than 50 mm), shape-dominated and scale-dominated rainfall. By first classifying areas that are either quite wet or quite dry, the shape and scale classification is only applied to those areas where rainfall amounts are in a critical range for supporting agriculture, and where drought conditions could have large effects on food security. This map (Figure 4) points out the scale-dominated regimes that are likely to have large variability in rainfall and result in irregular rainfall totals for January. Using the parameters in this way, it is possible to identify areas that are prone to dryness and drought-related issues. The countries highlighted in this map (Angola, Mozambique, Zambia and Zimbabwe) are all below the regional average in percentage of undernourished population (FAO, 2005b).

Establishing the connection between the estimated distribution parameters and the occurrence of extreme events is one of the primary objectives of this research. Generally speaking, areas with scale-dominated rainfall would experience more extreme and abnormal events. These areas have quite a range of rainfall amounts received, and need to have infrastructure and plans established to cope with extremely dry or wet conditions. Areas with shape-dominated rainfall, on the other hand, may experience more rain and may also have a larger absolute variance, but the large shape value indicates a relatively consistent accumulation from year to year. These generalizations allow for an interpretation of parameter values to be translated into conceptual regimes. Additionally, the parameters can provide an insight into

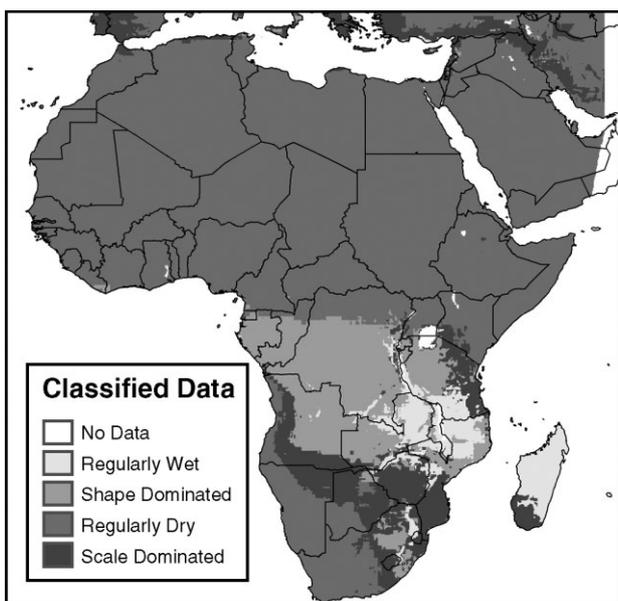


Figure 4. Results of decision-tree classification based on distribution parameters.

the suitability of an area in providing adequate rainfall for crops.

Defining 'extreme' is quite dependent on the particular use of the term. Farmers or decision makers concerned with crop health may be more concerned about the probability of no rainfall and dry events, as these are primary threats to productive agriculture. On the other hand, transportation managers may be more concerned with wet events and their ability to disrupt railway or automobile transportation. Finally, dam and hydroelectric engineers may be concerned with both wet and dry events to make sure that their waterways do not overflow their boundaries, while also keeping water in reserve to last through dry periods. The definition of extreme is different for each of these groups of people, but the combination of the probability of no rainfall and the gamma shape and scale should help affected individuals define the probability of catastrophe related to extreme rainfall.

KS test

The KS test was run using all points in the monthly subsampled data that contained more than one rainfall value in the climatological history. The results from this are displayed in Table I. A rejection level of 0.1 was used to test the rejection of the gamma distribution as suitable for parameterizing the CHARM record. For this data, including the points from all months, only 3.8% of points with at least two non-zero rainfall events in the 36-year history had a calculated p -value smaller than the 0.1 threshold (column 6 of Table I).

This test included points with very few rainfall values that tend to have a very high KS-statistic associated with them. To check how the inclusion of points with only a few rainfall values in the history negatively affected the overall statistics the test was performed again including only points that had rain in at least half of the history (a value of \hat{q} less than or equal to 0.5). Results for this are shown in Table II.

As expected, the average value for the KS-statistic (D_n) was reduced – by roughly a third – when including only points with over 17 values in the history. However, this reduction in the KS-statistic did not translate into a large increase in the average p -value. This is due to the fact that as more samples are included in the empirical dataset, the acceptable value of the KS-statistic becomes smaller. There was a notable improvement in the percentage of points with a p -value greater than 0.1. Of all points meeting the testing criteria, 98.5% had a p -value greater than 0.1, meaning that we do not reject the gamma distribution for 98.5% of the tested sites. If the more common threshold of 0.05 is used, the level of acceptance improves to 99.5% of all points with over 17 non-zero rainfall values.

This test shows that, overall, the gamma distribution appears to do an adequate job of approximating the historical rainfall distributions. To test how well the gamma distribution approximates the sampled data, it was compared with similar tests for the Weibull distribution. Results show that at the 0.1 rejection level, the

Table I. Results from KS test using all subset points with rainfall during the climatology.

Month	Mean p -value	Mean KS stat	# of points	Points <0.10	% >0.10	Points <0.05	% >0.05
Jan	0.6712	0.181	240 158	7 187	97.00	4586	98.09
Feb	0.6573	0.1783	234 527	10 277	95.61	7104	96.97
Mar	0.6618	0.1738	272 306	8 275	96.96	5040	98.15
Apr	0.6833	0.1786	283 612	8 657	96.94	6031	97.87
May	0.6529	0.1959	287 592	9 596	96.66	5934	97.94
Jun	0.631	0.2271	242 923	14 530	94.01	10 775	95.56
Jul	0.639	0.2067	219 988	10 383	95.28	6835	96.89
Aug	0.6305	0.2077	242 234	12 493	94.84	8219	96.61
Sep	0.6379	0.1984	264 040	11 809	95.52	7304	97.23
Oct	0.667	0.1833	284 573	7 758	97.27	4307	98.49
Nov	0.6822	0.1762	252 172	8 282	96.71	5327	97.89
Dec	0.6793	0.1803	241 737	6 982	97.11	4664	98.07

Table II. Results from KS test using only subset points with more than 17 rainfall values.

Month	Mean p -value	Mean KS stat	# of points	Points <0.10	% >0.10	Points <0.05	% >0.05
Jan	0.6941	0.1228	164 827	2178	98.67	615	99.63
Feb	0.6964	0.1226	176 800	2832	98.39	969	99.45
Mar	0.6845	0.1221	204 438	2956	98.55	970	99.53
Apr	0.7167	0.1176	207 333	1943	99.06	605	99.71
May	0.6916	0.1245	176 025	2226	98.73	563	99.68
Jun	0.6561	0.1299	134 481	2930	97.82	970	99.28
Jul	0.662	0.1282	130 712	2785	97.86	1018	99.22
Aug	0.6568	0.1288	146 066	3330	97.72	972	99.33
Sep	0.6628	0.1274	164 831	3492	97.88	1123	99.32
Oct	0.6881	0.124	191 325	2808	98.53	818	99.57
Nov	0.7151	0.1194	183 292	1713	99.06	491	99.73
Dec	0.7142	0.12	167 385	1596	99.04	509	99.70

performance of the gamma and Weibull parameter estimation techniques are quite comparable.

Given the similarities in the performance of the parameters in representing the samples, there are a few reasons to choose the gamma distribution over the Weibull. First, functions to define probabilities from gamma parameters already exist in many programming languages, while this is not the case for the Weibull distribution. Secondly, the MLEs are easier to calculate than the Weibull, reducing computation time when working with such a large grid. Finally, despite their complexity the gamma parameters are a bit more intuitive and interpretable, as well as more widely understood, for decision makers wishing to characterize their rainfall regimes. In the absence of a clearly superior distribution, these reasons support the decision of using the gamma distribution parameters to represent the rainfall probabilities for a given interval.

SUMMARY AND CONCLUSIONS

In summary, this study presented calculations of rainfall parameters for the gamma distribution using the

maximum likelihood estimates. The joint interpretation of monthly shape and scale parameters conveys the distribution of values in the modeled rainfall data at each location on the continent allowing the interpreter a qualitative assessment of the amount and stability of rainfall throughout the season. These parameters reflect the modeled rainfall, and as such also contain errors inherent in the modeled history.

The ability of the gamma distribution and parameter estimates to adequately fit the empirical distribution of values in the modeled history was tested using the KS goodness-of-fit test. This test showed that the gamma distribution and the estimated parameters could not be rejected as a suitable distribution for the CHARM historical data at a 0.10 confidence level for the majority of points on the continent. When considering only points that received rainfall for at least half the data history, the percentage of points that were acceptable at the 0.10 level increased slightly. This hypothesis testing indicates that the gamma distribution provides a reasonable description of the empirical rainfall probability distribution. The ability to represent the rainfall using the gamma

distribution parameters allows for interpretation of the parameter estimates as a compact summary of the full rainfall distribution.

Through analysis of the distribution parameters, it is possible to examine the likelihood of an area receiving rainfall amounts that would cause flooding, wash out dams or provide sufficient water to support crops. On the dry side of the distribution function, it is possible to understand quantitatively the range of possible drought scenarios that may occur at a location. The method implemented in this study is especially important in areas with spatially or temporally incomplete records, as is the case in Africa, because it allows for a more complete estimate of rainfall likelihood, given only a few records over a very large area.

The ultimate application of the information presented in this paper reveals the true value of this research. The types of applications could be generalized into two categories, those monitoring the existing conditions against the distribution, and those using the distribution to generate forecasts. One apparent monitoring application of the work presented here is the development of a continent-wide standardized precipitation index using near real-time rainfall estimates. In fact, a variety of monitoring tools based on the likelihood of the occurrence of observed events could be developed to estimate the impacts. An example of a forecast application would be to use the distribution information to develop rainfall scenarios that could drive agroclimatic models to assess typical crop conditions or end-of-season expectations for a region. These types of models could prepare decision makers for 'likely', 'best case', or 'worst case' scenarios to help them mobilize relief in a timely manner. Another example is to apply the distribution information at a watershed level to provide quantitative scenarios of rainfall volume and runoff granting information to hydrologic decision makers about how to manage available resources. Use of probability distributions in this way is certainly a future avenue of research.

Overall, this research shows how describing rainfall using a parametric distribution with parameter estimates fit to historical data increases the quality of the information about the rainfall history of an area. This research could prove valuable to a wide range of groups from scientists studying precipitation, to policy makers assessing forecast information, to local farmers estimating their crop yields.

REFERENCES

- Ananthakrishnan R, Soman MK. 1989. Statistical distribution of daily rainfall and its association with the coefficient of variation of rainfall series. *International Journal of Climatology* **9**: 485–500.
- Crutcher HL. 1975. A note on the possible misuse of the kolmogorov-smirnov test. *Journal of Applied Meteorology* **14**: 1600–1603.
- Food and Agriculture Organization. 2005a. *The State of Food and Agriculture 2005*. FAO: Rome.
- Food and Agriculture Organization. 2005b. *The State of Food Insecurity in the World 2005*. FAO: Rome.
- Funk C, Michaelsen J. 2004. A simplified diagnostic model of orographic rainfall for enhancing satellite-based rainfall estimates in data-poor regions. *Journal of Applied Meteorology* **43**: 1366–1378.
- Funk C, Michaelsen J, Verdin J, Artan G, Husak G, Senay G, Gadain H, Magadzire T. 2003. The collaborative historical African rainfall model: description and evaluation. *International Journal of Climatology* **23**: 47–66.
- Hayes MJ, Svoboda MD, Wilhite DA, Vanyarkho OV. 1999. Monitoring the 1996 drought using the standardized precipitation index. *Bulletin of the American Meteorological Society* **80**: 429–438.
- Ison NT, Feyerherm AM, Dean Bark L. 1971. Wet period precipitation and the gamma distribution. *Journal of Applied Meteorology* **10**: 658–665.
- Juras J. 1994. Some common features of probability distributions for precipitation. *Theoretical and Applied Climatology* **49**: 69–76.
- Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Leetmaa A, Reynolds B, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo KC, Ropelewski C, Wang J, Jenne R, Joseph D. 1996. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society* **77**: 437–471.
- Legates DR. 1991. An evaluation of procedures to estimate monthly precipitation probabilities. *Journal of Hydrology* **122**: 129–140.
- McKee TB, Doesken NJ, Kleist J. 1993. The relationship of drought frequency and duration to time scales. In *8th Conference on Applied Climatology*, 179–184.
- Öztürk A. 1981. On the study of a probability distribution for precipitation totals. *Journal of Applied Meteorology* **20**: 1499–1595.
- Siegel S. 1956. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill: New York.
- Thom HC. 1958. A note on the gamma distribution. *Monthly Weather Review* **86**: 117–122.
- Wilks D. 1990. Maximum likelihood estimation for the gamma distribution using data containing zeros. *Journal of Climate* **3**: 1495–1501.
- Wilks DS. 1995. *Statistical Methods in the Atmospheric Sciences: an Introduction*. Academic Press: San Diego, CA.
- Willmott CJ, Robeson SM. 1995. Climatologically aided interpolation (CAI) of terrestrial air temperature. *International Journal of Climatology* **15**: 221–229.
- Woolhiser DA. 1992. Modeling daily precipitation – progress and problems. In *Statistics in the Environmental & Earth Sciences*, Walden AT, Guttorp P. Halsted Press: London; 71–89.
- Wu H, Hayes MJ, Weiss A, Hu Q. 2001. An evaluation of the standardized precipitation index, the China-Z index and the statistical Z-Score. *International Journal of Climatology* **21**: 745–758.