Advances in Water Resources 94 (2016) 131–145

Contents lists available at ScienceDirect

Advances in Water Resources

journal homepage: www.elsevier.com/locate/advwatres

A global survey on the seasonal variation of the marginal distribution of daily precipitation

Simon Michael Papalexiou*, Demetris Koutsoyiannis

Department of Water Resources, Faculty of Civil Engineering, National Technical University of Athens, Heroon Polytechneiou 5, GR-157 80 Zographou, Greece

ARTICLE INFO

Article history:
Received 1 January 2016
Revised 7 May 2016
Accepted 10 May 2016
Available online 12 May 2016

Keywords:
Daily precipitation
Seasonal variation
Spatial variation
Marginal distribution
Generalized gamma distribution
Burr type XII distribution

ABSTRACT

To characterize the seasonal variation of the marginal distribution of daily precipitation, it is important to find which statistical characteristics of daily precipitation actually vary the most from month-to-month and which could be regarded to be invariant. Relevant to the latter issue is the question whether there is a single model capable to describe effectively the nonzero daily precipitation for every month worldwide. To study these questions we introduce and apply a novel test for seasonal variation (SV-Test) and explore the performance of two flexible distributions in a massive analysis of approximately 170,000 monthly daily precipitation records at more than 14,000 stations from all over the globe. The analysis indicates that: (a) the shape characteristics of the marginal distribution of daily precipitation, generally, vary over the months, (b) commonly used distributions such as the Exponential, Gamma, Weibull, Lognormal, and the Pareto, are incapable to describe "universally" the daily precipitation, (c) exponential-tail distributions like the Exponential, mixed Exponentials or the Gamma can severely underestimate the magnitude of extreme events and thus may be a wrong choice, and (d) the Burr type XII and the Generalized Gamma distributions are two good models, with the latter performing exceptionally well.

1. Introduction

"O, wind, if winter comes, can spring be far behind?"— P.B. Shelley

Most geophysical processes exhibit seasonal variation, which implies an underlying regular pattern, which potentially enables a degree of predictability, utilizing the periodic changes of the process's coarse behavior with time. This is exactly why it is important to correctly characterize the seasonal variability of geophysical processes. Among those, precipitation is one of the most important since it affects human lives significantly. Seasonality does not necessarily refer to the four standard seasons of the temperate zones, but it generally describes the within year variability. An effective scale to characterize seasonality is the monthly scale. Generally, planning and management of water resources systems, particularly those involving water supply (e.g. for irrigation) must take seasonality into account.

Precipitation may be represented as a stochastic process with two components: its marginal probability distribution and its dependence structure. We can reasonably expect these components to vary periodically if we study precipitation at any subannual time scale. Furthermore, it is rational to assume that the daily time scale is the finest time scale in which the seasonality could be studied without complications, because precipitation at subdaily scales may also be affected by earth's daily rotation (the daily cycle). In practice, estimating and trying to reproduce the statistical characteristics of precipitation on a daily basis can be a laborious task and, most importantly, can have questionable reliability as the estimation of the various characteristics will be based on small samples. For this reason, daily precipitation is typically studied and modeled on a monthly basis assuming that within a specific month its statistical characteristics remain essentially invariant. Consequently, the daily precipitation process can be decomposed into 12 different processes with fixed month-to-month correlations and fixed monthly marginal distribution. Here we are not concerned with the autocorrelation structure but we focus on the monthly variation of the marginal distribution of the daily precipitation.

The marginal distribution of daily precipitation belongs to the so-called mixed type distributions and comprises two parts: a discrete part describing the probability dry and mathematically expressed as a probability mass concentrated at zero, and a continuous part spread over the positive real numbers describing probabilistically the amount or the intensity of nonzero precipitation. The probability dry, in general, can be easily assessed from empirical data as the ratio of the number of dry days over the total number of days, while the continuous part is usually modeled by a parametric continuous distribution fitted to nonzero values. Yet this distribution is not unique and in practice, as a literature

* Corresponding author.
E-mail address: smp@itia.ntua.gr (S.M. Papalexiou).

http://dx.doi.org/10.1016/j.advwatres.2016.05.005
0309-1708/

review reveals, various distributions have been used for the nonzero daily precipitation. For example the Exponential distribution (e.g., Smith and Schreiber, 1974, Todorovic and Woolhiser, 1975), mixed Exponentials (e.g., Woolhiser and Roldán, 1982, Wilks, 1998, Wilks, 1999), the Gamma distribution (e.g., Buishand, 1978, Bruhn et al., 1980, Geng et al., 1986), the Weibull distribution (e.g., Swift and Schreuder, 1981, Wilson and Toumi, 2005), the Lognormal distribution (e.g., Swift and Schreuder, 1981, Biondini, 1976), mixed Lognormals (Shimizu, 1993), power-type distributions like the two-, three- and four-parameter Kappa distributions (Mielke Jr, 1973, Mielke Jr and Johnson, 1973, Hosking, 1994, Park et al., 2009), generalized Beta distributions (Mielke Jr and Johnson, 1974), as well as the Generalized Pareto (e.g., Fitzgerald, 1989) for peaks over threshold, and probably many more.

A question that can be raised based on the aforementioned studies and on many more is whether or not all of these distributions, some completely different with each other in structure, are indeed suitable for describing the probability of nonzero daily precipitation or they have prevailed and become popular for reasons such as simplicity. Additionally, most of these studies are of local character, i.e., they are based on the analysis of a limited number of precipitation records and from specific areas of the world. The exceptions are very few, e.g. in a study by Papalexiou and Koutsoyiannis, (2012) daily precipitation was analyzed in more than 10,000 stations worldwide. In practice, in most cases precipitation in modeled using exponential-type distributions like the Exponential distribution, the Gamma or mixed Exponentials. These, however, might not be adequate if the actual distribution of nonzero precipitation has a heavier tail than those light tail distributions and consequently may severely underestimate the magnitude and the frequency of extreme events. Actually, two recent studies (Papalexiou and Koutsoyiannis, 2013, Papalexiou et al., 2013), where daily precipitation extremes were analyzed in more than 15,000 stations worldwide, revealed that most of the records cannot be described by exponential-tail distributions but rather by distributions with heavier tails.

In this study the seasonal variation of the marginal distribution function of daily precipitation is analyzed to find which statistical characteristics of daily precipitation actually vary the most from month-to-month and which could be regarded to be invariant. Relevant to the latter issue is the question whether there is a single model capable to describe effectively the nonzero daily precipitation for every month and at every area of the world. Obviously these questions cannot be answered by local analyses. Therefore, here we perform a massive analysis approximately at 170,000 monthly daily precipitation records from more than 14,000 stations from all over the globe.

2. The data

The original database we use here is the Global Historical Climatology Network-Daily database (version 2.60, www.ncdc.noaa.gov/oa/climate/ghcn-daily) which comprises thousands of daily precipitation records from stations all around the globe. Nevertheless, we use only a part of these records as many of them are very short in length, contain a large percentage of missing values, or have values of questionable accuracy which are assigned with various quality flags (details on quality flags can be found in the website given above). For these reasons and in order to create a robust subset of records with ensured quality we chose only those having: (a) record length longer than 50 years, (b) missing values less than 20% and, (c) values assigned with quality flags less than 0.1%. As an additional measure to ensure the quality of the data we deleted all values assigned with flags "G" (failed gap check) or "X" (failed bounds check) as these flags are used for unrealistically large values. Fortunately, only 594 records in total had such values and typically no more than one or two values per record. The resulting subset comprises 15,137 stations.

Although this study concerns the monthly daily precipitation we analyze also the daily precipitation of all months as in some cases, especially for design purposes, we are not interested about the month that an event occurs but just on its exceedance probability or else on its return period. In this case monthly daily values can be merged and treated as represented by a single random variable (note that the term "daily precipitation" refers to daily precipitation values of all months while the term "monthly daily precipitation" refers to the daily precipitation values of individual months). From each station we formed 13 different records, one for all daily values and 12 for the monthly daily values, resulting in a total of 196,781 different records. Nevertheless, some months for stations located in very dry areas have very few nonzero precipitation values or even none so that estimation of the various important statistics would be highly uncertain or even impossible (e.g., estimation of L-skewness needs at least three values). To overcome this problem we constrained the minimum sample size of monthly nonzero precipitation values; so among the 15,137 records initially chosen we finally selected those having at least 20 nonzero values for each month resulting in a total of 14,157 stations and consequently 169,884 monthly daily records were formed. The locations of these stations and their corresponding lengths in years are given in the map of Fig. 1. Note that in some areas the map cannot provide the clear picture of the record length distribution. For example in the USA, the network of stations is very dense and inevitably points overlap, so that, below the layer of points representing high record lengths, other points exists representing smaller records lengths.

3. Seasonal variation

3.1. Statistics studied

To assess the seasonal variation of daily precipitation we study representative statistics of the marginal distribution on a monthly basis. Additionally, in order for the study to be more complete as well as for comparison purposes we estimated these statistics for the daily precipitation values of all months too (indicated with "All" in the figures). Particularly, we studied: (a) the probability dry, (b) the mean value, (c) the L-variation, and (d) the L-skewness. The probability dry expresses the discrete part of the marginal distribution and is simply estimated as the ratio of dry days to total days. The latter three are statistics for the continuous part of the marginal distribution describing the nonzero precipitation, which are calculated using only nonzero precipitation values.

The mean value of nonzero precipitation is a classical measure of central tendency while L-variation $\tau_2 = \lambda_2/\lambda_1$ and L-skewness $\tau_3 = \lambda_3/\lambda_2$, defined as ratios of L-moments λ_i (Hosking, 1990), are dimensionless measures of the distributional shape. L-ratios are preferable over ratios based on the classical moments like the coefficients of skewness and kurtosis as they exhibit better statistical properties, e.g., they are more robust (see e.g., Hosking, 1992). Additionally, L-kurtosis (defined as $\tau_4 = \lambda_4/\lambda_2$) is also commonly used as a measure of shape, yet for positive random variables L-variation is well defined and actually is more robust and more convenient as it is bounded in [0,1]. Usually, L-variation or even the classical coefficient of variation (defined as the ratio of standard deviation to the mean value) are interpreted as standardized measures of variance; indeed, they express, respectively, the value of the second L-moment λ_2 and the value of the standard deviation of a distribution having mean value equal to 1. Yet for positive random variables, where actually these coefficients are meaningful, both depend on the distribution's shape parameters only or are constants