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

Simon Michael Papalexiou and Demetris Koutsoyiannis

Department of Water Resources, Faculty of Civil Engineering, National Technical University of Athens, Heroon Polytechneiou 5, GR-157 80 Zographou, Greece (smp@itia.ntua.gr)

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.

Keywords: daily precipitation, seasonal variation, spatial variation, marginal distribution, Generalized Gamma distribution, Burr Type XII distribution

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 review reveals, various distributions have been used for the nonzero daily precipitation. For example the Exponential distribution [e.g., 1,2], mixed Exponentials [e.g., 3–5], the Gamma distribution [e.g., 6–8], the Weibull distribution [e.g., 9,10], the Lognormal distribution [e.g., 9,11], mixed Lognormals [12], power-type distributions like the two-, three- and four-parameter Kappa distributions [13–16], generalized Beta distributions [17], as well as the Generalized Pareto [e.g., 18] 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 non-zero 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 [19] 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 [20,21], 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 Figure 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 exist 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 [22], 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., 23]. 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 if the distribution does not have shape parameters, and thus, they are essentially measures of distributional shape.

As already noted, we anticipate from our experience the probability dry to vary over the months in most areas of the world. Additionally, it may seem obvious that the monthly mean value of daily precipitation (including zero values) will vary too as it is directly related to probability dry, e.g., a larger number of rainy days on average in a month logically will increase the monthly mean (estimated as the record's total monthly precipitation divided by the total number of month's days). However, it is not that evident that the mean value of the monthly nonzero daily precipitation (estimated as the record's total monthly precipitation divided by the total number of the month's rainy days) will vary over the months (during rainy days it could be possible to rain on average the same amount irrespective of the month). Finally, our perception on precipitation may lead us to assume that extreme precipitation varies with season, e.g., it is well-known that specific weather mechanisms, responsible for extreme precipitation, are linked with specific seasons. Consequently, this may imply that the shape characteristics of precipitation distribution change over seasons, as the distribution's shape, particularly the right tail, controls the frequency and the magnitude of extreme events. Yet this assumption may be false as extreme precipitation may emerge by a change in the scale or else in the variance of precipitation and not necessarily by a change in its shape characteristics. For these reasons whether or not the distributional shape characteristics vary with season needs to be investigated and verified.

3.2 Variation in the hemispheres

Northern Hemisphere (NH) and Southern Hemisphere (SH) have opposite seasons and thus, it is reasonable to assume that natural processes under seasonal variation exhibit different behavior between the two hemispheres. This may be generally valid, especially for processes like the surface temperature, yet precipitation is a more complex process that may be affected more by regional climate conditions. For example, the celebrated Köppen climate classification [see e.g., 24,25], which classifies climate according to the annual and monthly average temperature and precipitation, defines several different types and subtypes of climate for each hemisphere. Thus, different precipitation patterns may appear even in adjacent areas of the same hemisphere.

Nevertheless, a first coarse approach that could provide a general picture is to present the seasonal variation of the statistics by hemisphere. Among the 14,157 stations analyzed, 8447 belong in the NH and 5710 in the SH. The aforementioned statistics, i.e., the probability dry, mean value, L-variation and L-skewness, were calculated for the monthly daily precipitation of each station; their averages and standard deviations are given, for each hemisphere and additionally for the whole globe, in Table 1. Furthermore, a better picture is provided by the box plots given in Figure 2 which present these statistics on a monthly basis and for each hemisphere. The left (red) box plots are for the NH while the right (gray) are for the SH while the box plot's inner lower and upper fences that define the box indicate, respectively, the 25% and 75% empirical quantile points and thus define the empirical interquartile range (IQR) or the 50% of the central values. The line within the box indicates the median, while the lower and upper fences of the whiskers indicate, respectively, the 5% and 95% empirical quantile points or else they define the 90% empirical confidence interval (ECI) of the studied statistics. It should be clear that results presented for each hemisphere express the average and standard deviation values of the stations analyzed in each hemisphere and may not be representative values for the whole hemisphere (especially in the SH where stations are situated in few areas). Estimation of representative hemisphere values, if possible, would demand spatial integration which is out of the scope of this study.

As we see in Figure 2, the probability dry in NH exhibits the typical behavior we have in our minds for NH, i.e., dry summer months and wet winter months. Particularly, if we focus on the median of each box plot it exhibits a sinusoidal-like variation, so it seems that most stations in NH have this pattern. Surprisingly, the corresponding pattern in SH is not clear at all; if we focus on the median, although it resembles a sinusoidal-like function, clearly, it is not the familiar and the anticipated one as it has three “local” peaks, i.e., in January, April and August. We also note that the IQR seems to vary irregularly and does not follow the variation of the median. Of course, this does not imply the absence of seasonality in probability dry in the SH, as this result can easily emerge if we assume several different patterns for the studied stations. Also, it is interesting that the variation of the median in both hemispheres is not very large, especially in the SH, yet the range of the 90% ECI is very wide expressing the large variation of probability dry around the world.

The mean value of the nonzero precipitation in both hemispheres, as Figure 2 shows, exhibits a clear seasonal pattern, which reminds that of the surface temperature. Specifically, NH and SH show essentially a contrasting behavior to each other, yet in terms of seasons the behavior is the same, i.e., the warm months in both hemispheres are those with the highest average nonzero daily precipitation. This behavior though is not in full correspondence as in NH the minimum and the maximum mean values (comparing the medians) are, respectively, in January and in September, while the corresponding values in the SH are observed, respectively, in August and in February. Remarkably, for the NH the average nonzero daily precipitation pattern is in contrast with probability dry implying greater precipitation depths in rainy days of dry months than of wet months. Yet this is not absolutely precise as the driest moths are from June to August while those with the highest average of nonzero daily precipitation are from July to September; additionally, the lowest value in probability dry is in July while the peak average value is in September. This contrast seems not to be valid for the SH as the probability dry exhibits an irregular pattern.

Figure 2 also reveals a marked monthly variation pattern for L-variation and L-skewness. Similarly to the average of nonzero daily precipitation, both statistics exhibit a contrasting behavior between the two hemispheres; but again, comparing the medians, high and low values are observed, respectively, at warm and cold months. A comparison between the two shape statistics shows that L-variation and L-skewness in SH show an almost identical pattern with the only difference being in the lowest value which is observed one month later for L-skewness. Additionally, L-variation in NH takes its lower values around February while L-skewness around April. Generally, the monthly variation of both statistics (based on their medians) is small, i.e., in both hemispheres L-variation and L-skewness range, respectively, from 0.55 to 0.6 and from 0.42 to 0.47. However, the IQR or the 90% ECI is much wider in the SH compared to NH. Comparing the shape statistics with the mean value of daily precipitation, we note an agreement in the general pattern in SH, while in NH especially for L-skewness the difference in the patterns is significant.

3.3 A simple test to identify seasonal variation

All previous comparisons based on the monthly box plots of the statistics indicate clear seasonal variation patterns; a surprising exception is the probability dry of the SH. Nevertheless, both the

IQR and the 90% ECI of all those statistics are much wider allowing at least theoretically a portion of the stations studied to have different patterns than the characteristic one indicated by the medians in Figure 2.

As mentioned, we intuitively anticipate some characteristics of daily precipitation like the probability dry to vary with season, yet this it is not self-evident, e.g., for distributional shape measures like L-variation and L-skewness. When dealing with a small number of records it is relatively easy to assess if a statistic varies with season using simple means, e.g., a plot of the statistic *vs.* month would reveal the variation pattern. Yet when dealing with thousands of stations, an “eyeball” technique would be insufficient or even subjective. For this reason we form here a simple test to assess and quantify the seasonal variation of the various statistics we investigate.

Seasonal variation evokes sinusoidal-like functions; however, even if a statistic is expected to obey a sinusoidal-like law, its sample counterpart may deviate significantly from the anticipated law due to sample variability commonly caused either by sampling uncertainty, particularly for small samples, or by non-robust estimators, or even from local weather characteristics modifying the expected behavior in some months. This implies that a precise sinusoidal variation may not be common to observe and thus a test based on these characteristics would be inflexible and probably with doubtful efficacy. For this reason, we propose here a non-parametric test allowing for the statistic under investigation to deviate from the exact sinusoidal form.

The seasonal variation test (SV-Test) is described in the following steps: (a) the desired statistic is calculated for each month, (b) the numbers 1 and −1 are assigned, respectively, to monthly values smaller and larger than the median of all months, (c) this sequence is rotated until the first and the last value have different signs, (d) this sequence is split into sub-sequences consisting of identical-value runs (SIVR), (e) the number of SIVR is calculated.

Note that step (c) is necessary to simplify the test and estimate less benchmark values by the Monte Carlo process described in sequence. Particularly, if step (c) is not applied then two cases should be studied, i.e., one when the sign between the first and last value differs, and one when it is the same. Given that six values will equal 1 and six −1 (that emerges by the definition of the median value) it can be proven that if the sign differs then the number of feasible SIVR that a sequence consisting 1 and −1 can be split is 2, 4, 6, 8, 10 or 12 if all values are alternating. In the other case an odd number of SIVR would emerge, i.e., 3, 5, 7, 9 or 11. Also, step (c) ensures that the resulting number of SIVR is the minimum, e.g., a sequence starting and ending with the same sign having 11 SIVR if it is rotated in order the first and last sign to differ it will have 10 SIVR.

The resulting number of SIVR quantifies seasonality. If the considered statistic exhibits a sinusoidal-like seasonal variation the SV-Test will result exactly in two SIVR. Figure 3 depicts an explanatory sketch of the SV-Test showing the monthly values of a statistic after rotation so that the first and the last value are in opposite sides of the median; even though the statistic does not resemble exactly a sinusoidal law, the application of the test results in two SIVR revealing the seasonality that is visually apparent. We could also expect that four SIVR still reveal seasonal variation as they could easily emerge if the statistic's sample estimates are sensitive, e.g., if the December's value in the graph of Figure 3 was above the median, then four SIVR would result. It seems reasonable to assume that a larger resulting number of SIVR indicates random variation or a variation that does not resemble the "familiar" seasonal variation.

One could argue that the previous interpretation of the resulting number of SIVR is subjective, e.g., it could be assumed that two or four SIVR could easily emerge even if there is no seasonal variation due to randomness. Thus, in order to make the SV-Test complete we need benchmark values for reference and comparison. The idea is to find the probability for each feasible number of SIVR to emerge in the case where the variation of a statistic is random. Theoretically, this problem can be solved analytically using combinatorics, yet it is not that easy; in contrast a Monte Carlo approach can easily provide the answer. In this direction, we apply a Monte Carlo simulation summarized in three simple steps: (a) we generate 10^6 samples consisting of 12 random numbers each, (b) we apply the SV-Test to estimate the resulting number of SIVR for each sample, and (c) we calculate the probability for each feasible number of SIVR as the ratio of the times that this number of SIVR emerged to total number of samples (10^6).

The results are graphically depicted in Figure 4 where the first number above the bars indicates the probability for a specific SIVR number to occur and the second number above the bars indicates the cumulative probability, e.g., the probability for up to four SIVR to occur is 17.6%. Accordingly, if a statistic varies randomly the probability for two SIVR is only 1.3% and for four is 16.3%, while the most probable numbers of SIVR are six and eight with probabilities 43.3% and 32.5%, respectively. This implies that if the studied statistic does not exhibit seasonal variation then application of the test will result in more than two SIVR with probability 98.7% and in more than four SIVR with probability 82.4%, and thus, we can safely assume that not only two but also four SIVR indicate seasonal variation.

3.4 Application of the test

We applied the SV-Test for each station and for the four aforementioned statistics with the results presented in Figure 5. The SV-Test verifies, as we see in Figure 5a, that indeed probability dry exhibits seasonal variation with 64.1% of the stations resulting in two SIVR and with only 4.9% of the stations resulting in more than four SIVR indicating random variation. Similar results are obtained for the mean value of the nonzero daily precipitation, given in Figure 5b, with only 8.3% of the stations resulting in more than four SIVR.

The results of the SV-Test regarding the shape characteristics of the nonzero daily precipitation, i.e., the L-variation and the L-skewness are depicted, respectively, in Figure 5c and Figure 5d. The first we note is that the profile of the two graphs is completely different from the “benchmark” graph describing the random case in Figure 3; however, the results are not as clear as for the probability dry or for the mean value case. We see that the most common SIVR number is four, both for L-variation and for L-skewness, with 36.9% and 34.5%, respectively. Nevertheless, two or four SIVR (numbers indicating seasonal variation) emerge at 66.2% of stations for L-variation and at 54.5% of stations for L-skewness, while the corresponding value for the random case is much smaller, i.e., 17.6%. Additionally, two SIVR are observed in 29.3% and 19.7% of the records for L-variation and L-skewness, respectively. These percentages are much larger than 1.3%, which corresponds to the random case. Finally, the seasonality signal is it is much stronger for L-variation than for L-skewness, a difference that may attributed in the fact that estimation of L-variation is more robust than L-skewness.

3.5 Why and how much statistics vary?

Studying the statistics by hemisphere as well as the results of the SV-Test revealed that seasonal variation occurs not only in probability dry and in the mean value of nonzero precipitation but also in the shape characteristics. This implies that the marginal distribution varies over the months, yet the mechanism of this variation is not clear. Particularly, different aspects of the precipitation process are interrelated. For example, the distributional shape variation may be affected by seasonal variation of the average storm duration. To clarify by an example, let us consider the random variables *X* and *Y* representing, respectively, the amount of nonzero precipitation at the daily and at a much finer time scale, e.g., the one-minute scale, and let us assume that the marginal distribution of *Y* does not have seasonal variation; then the distribution function of *X* emerges by the *n*-term sum of *Y* variables where *n* corresponds to the storm duration in minutes in that particular day. Cleary, if the average storm duration varies per month, then the "average" *n*-term sum will vary too and hence the distribution of *X*. This issue raised can only be answered by an analysis of fine temporal scale data which is not the subject of this particular study.

In order to quantify the seasonal variation of the studied statistics per station we define four difference measures relative to the statistic's average value of all months. These measures are illustrated in the sketch of Figure 6 depicting the monthly variation of a statistic. Particularly, we define the *i*-th monthly difference $D_i = V_i - \mu$ as the difference between the *i*-th month statistic's value V_i and the average of all V_i denoted as μ. Negative differences (blue lines in the graph) are denoted with D_{N} and their average with $\bar{D}_{\mathrm{N}}$; likewise, D_{P} denotes positive differences (red lines in the graph) and $\bar{D}_{\mathrm{P}}$ denotes their average. Additionally, $D_{\min}$ and $D_{\max}$ denote, respectively, the minimum and the maximum difference with reference to μ. Note that this analysis in performed for each individual station and does not provide any comparison between different stations.

The difference measures $\bar{D}_{\mathrm{N}}$, $\bar{D}_{\mathrm{P}}$, $D_{\min}$ and $D_{\max}$ are calculated in terms of percentage change (PC) in respect to the average μ, i.e., $\mathrm{PC} = 100\, D / \mu$ with *D* being any of the four difference measures. The first two measures can be interpreted as the "expected" or the average negative or positive percentage change in reference to the monthly average while the latter two indicate the minimum or maximum percentage change in reference to the monthly average. We calculated the percentage change of these measures for each station and for the four statistics studied. The results are given in Figure 7 in the form of box plots (note that the PC of the negative differences $\bar{D}_{\mathrm{N}}$ and D_{min} is given in absolute values for better presentation).

A first look in the box plots indicates that the largest monthly variation is observed in the mean value of the nonzero precipitation, followed by the probability dry, next by L-skewness and last by L-variation exhibiting the lowest variability. Particularly, the IQR of the nonzero precipitation mean value, which represents the 50% of the central values, for D_{min} and D_{max} ranges, respectively, from −45.2% to −22.8% and from 25.5% to 50.6%; these values indicate a large variability around the average. These ranges are lower for the probability dry where the IQR of D_{min} and D_{max} ranges, respectively, from −24.3% to −9.2%) and from 8.2% to 19.2%. Regarding L-skewness we observe that 75% of the records have percentage change of D_{min} and D_{max} less than −17.7% and 20.4%, respectively, while the corresponding percentages for the L-variation are −9.5% and 10.7%. Comparing the box plots of the distributional shape measures, i.e., the L-variation and L-skewness, with the box plots of the probability dry and of the mean value we observe that in the first two cases $\bar{D}_{\mathrm{N}}$ and $\bar{D}_{\mathrm{P}}$ vary at a lower level relative to D_{min} and D_{max} than in the former two cases. This may indicate that the "expected" difference from the monthly average, expressed by $\bar{D}_{\mathrm{N}}$ and $\bar{D}_{\mathrm{P}}$, for L-variation and L-skewness for most of the months is "small"; yet the "extreme" differences, expressed by D_{min} and D_{max}, are relatively large; or else, this indicates that the marginal distribution of nonzero daily precipitation for most of the months does not vary much in terms of shape.

4. In search for the “universal” precipitation model

4.1 Candidate models

The shape characteristics of nonzero daily precipitation, as empirical evidence suggests, vary not only with location but also by month; this implies that the consistent probabilistic modeling of nonzero daily precipitation demands different models for different areas and possibly for different months. So it would be of paramount importance if a single parametric distribution can be used for nonzero daily precipitation for all months and for the whole world. The fact that distributional shape varies excludes, in principle, distributions with fixed shape, thus favoring those with great shape flexibility. Additionally, we deem that a competitive model should also be physically consistent with precipitation, i.e., defined in the positive real axis, and if possible having a theoretical basis. In this direction, in a previous study [19] we used the principle of maximum entropy to derive consistent distributions for geophysical random variables. These entropy derived distribution were tested in their ability to describe the nonzero daily precipitation (but not in a monthly basis) using more than 10,000 stations with very good results.

The distributions derived in the aforementioned study, and also used here are the Burr type XII distribution (BrXII) [26,27] and the Generalized Gamma distribution (GG) [28]. Their probability density functions are given, respectively, by

$$f_{\mathrm{BrXII}}(x) = \frac{1}{\beta}\left(\frac{x}{\beta}\right)^{\gamma_1 - 1}\left(1 + \gamma_2\left(\frac{x}{\beta}\right)^{\gamma_1}\right)^{-\frac{1}{\gamma_1\gamma_2} - 1} \qquad x \geq 0 \tag{0}$$

$$f_{\mathrm{GG}}(x) = \frac{\gamma_2}{\beta\,\Gamma\left(\gamma_1/\gamma_2\right)}\left(\frac{x}{\beta}\right)^{\gamma_1 - 1}\exp\left(-\left(\frac{x}{\beta}\right)^{\gamma_2}\right) \qquad x \geq 0 \tag{0}$$

Note that the parameterization we use here for the BrXII is different from the most typical found in the literature; first, it clearly shows its asymptotic behavior (for $\gamma_2 \to 0$ the Weibull distribution emerges) and second, the two shape parameters are directly related to each of the distribution tails (left and right). Regarding the parameterization of GG distribution we mention that other forms also exist but this is one of the commonly used.

Both distributions are very flexible, each comprising one scale parameter $\beta > 0$, and two shape parameters. The shape parameter $\gamma_1 > 0$ controls the behavior of the left tail, i.e., for $\gamma_1 < 1$ the distributions are J-shaped while for $\gamma_1 > 1$ they are bell-shaped; the parameter $\gamma_2 > 0$ controls the asymptotic behavior of the right tail, i.e., the "heaviness" of tail and thus the frequency and the magnitude of extreme events. It is noted that although these two distributions have a structural similarity in terms of their parameters, in principle, they differ, i.e., the BrXII distribution is a power-type distribution having finite moments up to order $1/\gamma_2$ while the GG distribution is of exponential form with all of its moments finite. Some well-known special cases worth mentioning for the BrXII distribution are the Pareto type II and the Weibull distributions (limiting case), while for the GG distribution, special cases are the Weibull, the Gamma and the Exponential distributions.

4.2 A first approach based on L-moments

There are some useful graphical tools, especially when dealing with a large number of records, which help to provide an overall and general picture of the studied variable from a statistical point-of-view. Such a tool for identifying suitable distributions for the variable under investigation is the L-moments ratio diagram [see e.g., 29,30]. Essentially, this diagram provides a comparison between observed statistics calculated from the records and the theoretical ones emerging by the distribution under investigation. Practically, any pair of L-ratios could be used to form an L-ratio diagram; yet the most common pairs are the L-skewness *vs*. L-variation or the L-kurtosis *vs*. L-skewness, with the latter being more popular in the literature as L-variation is not well defined for some distributions, e.g., for distributions with mean value zero or negative. Nevertheless, as noted, L-variation is well defined for positive random variables and is more robust than L-kurtosis.

L-ratios as functions of the distribution's shape parameters are essentially measures of shape. Thus, in an L-ratio diagram a distribution with none, one or two shape parameters forms, respectively, a point, a line or an area. Consequently, the aforementioned distributions, in any L-ratio diagram, form an area (denoted as L-area) whose extent is finite (does not cover the entire plane). Here we use the L-skewness *vs*. L-variation diagram aiming to form the theoretical L-area of the BrXII and the GG distributions and calculate the percentage of the observed L-points that lie within the L-area of each distribution and for each month. An observed point that lies within the distribution's theoretical L-area implies that specific parameter values exist so the distribution can reproduce the first three L-moments. Practically, the theoretical L-area of a distribution is formed using equations of τ_2 and τ_3. Unfortunately, analytical L-moment expressions for the GG distribution do not exist; exception is the first L-moment (identical with the mean value) and is given by

$$\lambda_1 = \beta\,\Gamma\left(\frac{1+\gamma_1}{\gamma_2}\right) / \,\Gamma\left(\frac{\gamma_1}{\gamma_2}\right) \tag{0}$$

where $\Gamma\left(a\right) = \int_0^\infty t^{a-1} \exp(-t)\mathrm{d}t$ is the Gamma function. In contrast, for the BrXII distribution, solving the L-moments definition integrals [see e.g., ,22], we found the following expressions:

$$\lambda_1 = \frac{\beta\gamma_2^{-1/\gamma_1}}{\gamma_1} \mathrm{B}\left(\frac{1}{\gamma_1}, \frac{1-\gamma_2}{\gamma_1\gamma_2}\right) \tag{0}$$

$$\tau_2 = 1 - \mathrm{B}\left(\frac{1}{\gamma_1}, \frac{2-\gamma_2}{\gamma_1\gamma_2}\right) / \mathrm{B}\left(\frac{1}{\gamma_1}, \frac{1-\gamma_2}{\gamma_1\gamma_2}\right) \tag{0}$$

$$\tau_3 = 1 - 2\frac{\mathrm{B}\left(\frac{1}{\gamma_1}, \frac{2-\gamma_2}{\gamma_1\gamma_2}\right) - \mathrm{B}\left(\frac{1}{\gamma_1}, \frac{3-\gamma_2}{\gamma_1\gamma_2}\right)}{\mathrm{B}\left(\frac{1}{\gamma_1}, \frac{1-\gamma_2}{\gamma_1\gamma_2}\right) - \mathrm{B}\left(\frac{1}{\gamma_1}, \frac{2-\gamma_2}{\gamma_1\gamma_2}\right)} \tag{0}$$

where $\mathrm{B}(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\,\mathrm{d}t$ is the Beta function. The two parametric equations $\tau_i = g_i(\gamma_1, \gamma_2)$ given in equations (0) and (0) can be used to implicitly determine the L-area. Functions of this form, and in this particular case, can be easily plotted by fixing one parameter to a specific value, varying the other in a dense grid and plotting the resulting (τ_2, τ_3) points. The method for determining the theoretical L-area covered by the GG distribution is exactly the same, with the only difference that (τ_2, τ_3) points are calculated by the numerical integration of the L-moments integrals.

The theoretical BrXII and GG L-areas are depicted in Figure 8, with several fixed-value parameter lines also plotted. For the BrXII distribution values ranging from 1 to 10 (lower bound) denote fixed γ_1 parameter values while those ranging from 0.1 to 0.9 (upper bound) denote fixed γ_2 parameter values. Similarly, for the GG distribution values ranging from 0.5 to 6 (lower bound) denote fixed γ_1 parameter values while those ranging from 0.5 to 10 (within the area) denote fixed γ_2 parameter values. The observed L-points of the nonzero daily precipitation for the month of January are also shown in Figure 8, superimposed over the L-areas (graphs for individual months as well as for the nonzero daily precipitation of all months are given as supplementary material). At each plot empirical points are colored in three ways; the red-colored points lie outside the area; the dark-colored indicate a Bell-shaped distribution; the light-colored indicate a J-shaped distribution. Interestingly, the GG and the BrXII distributions are complementary in the sense that the observed L-points not belonging to one's area belong to the other's, implying that just these two distributions can describe all records analyzed here. Note that both distributions are special cases of the Generalized Beta of the second kind distribution [see e.g., 17,19], but this distribution is more complicated as it comprises one scale and three shape parameters.

Particularly, Figure 9 shows the estimated percentages of the observed L-points of monthly daily precipitation lying within the area. We also display the percentages of J- and Bell-shaped distributions that would emerge if the distributions were actually fitted. It is apparent that both distributions, especially the GG distribution, perform very well. For example, the GG distribution describes 99.2% of the observed L-points for the values of all months, while the lowest percentage, observed in January, remains very high, i.e., 94.2%. The BrXII distribution also performs well by managing to describe 90.0% of the observed L-points for the values of all months and with its lowest percentage observed in May with 81.0%. We note that the actual percentages of the observed points that lie within the theoretical areas are expected to be even higher if larger samples were available. Clearly, the variability of the statistics decreases with increasing sample size and thus many points that lie outside the area actually would not if the sample was larger. Actually, this is the reason why the percentage of the observed L-points for the values of all months is higher than those of individual months. Finally, it may seem peculiar that the percentages of J-shaped GG distributions are significantly lower (almost half) compared to those of the BrXII distributions. This implies that for the same record a J- and a Bell-shaped distribution may be fitted equally well in terms of L-moments. Note that a density function $f(x)$

is called J-shaped if the value of $f(x)$ at its lower bound (zero for positive random variables) is the maximum, i.e., $f(0)=\max\left(f(x)\right)$; otherwise, the distribution is called Bell-shaped. This simple criterion may however be meaningless in several practical situations, e.g., two GG distributions with γ_1 values a little less and a little more than 1 would be characterized, respectively, as J- and Bell-shaped, yet apart from this difference they are almost identical.

The previous analysis gave a clear indication that both the GG and the BrXII distributions are very good models for describing precipitation. Yet an important and more specific question that naturally arises is if a single distribution can be used to describe all months within the same station; in order to answer this question an analysis by record has to be performed. To clarify, each record has 12 L-points, one for each month, so the idea is to estimate the number of monthly L-points per station that lie within the theoretical L-area. For example, if all monthly points of a station lie within the distribution's area, then this distribution could be used for all months in this particular station. The results are shown in Figure 10. Evidently, in this test the GG distribution performs much better than the BrXII, as it can be used as an all-month model for 78.8% of the stations, a percentage almost double than the corresponding one to the BrXII distribution which is 43.2%. Additionally, the percentage of record in which the GG distribution is suitable for more than ten months is very high, i.e., 95.6% while the corresponding one for the BrXII it has significantly increased to 69.5%.

4.3 The actual fitting

The previous analysis showed that both distributions can describe a very large percentage of the records in terms of the first three L-moments. Additionally, it is very important to study the actual values of the shape parameters, especially of the parameter γ_2 as it controls the extreme behavior. As noted though, the GG distribution does not have analytical L-moments equations while in the BrXII case, where analytical formulas exist, the resulting system of equations between theoretical and sample estimates can only be solved numerically. So it is clear that explicit functions, easily applicable, of the form $\theta = g(\lambda_1,\tau_2,\tau_3)$ that relate any of the distribution's parameter θ with the first three L-moments measures cannot be formed.

In order to create a fitting method for both distributions that is based on L-moments and is accurate and fast to apply, we approach the problem inspired by the way engineers and statisticians used to practice in the past (or even at present) using the "good-old" graphical tools (e.g., nomograms). For example, the shape parameters γ_1 and γ_2 can be approximately estimated by placing an observed (τ_2,τ_3) point within the L-ratio diagram in Figure 8 and do an "eyeball" linear regression using the nearest fixed-value parameter lines surrounding the observed point. Essentially, our approach is an accurate and computerized version of this technique, i.e., the algorithmic "translation" of a (τ_2,τ_3) point to a (γ_1,γ_2) point. The basic idea is to "replace" the initial functions of L-variation and L-skewness, which are highly nonlinear and without analytical expressions in the GG case, with simple linear interpolation functions that can be more easily handled. First, we calculate $\tau_2 = g_2(\gamma_1,\gamma_2)$ and $\tau_3 = g_3(\gamma_1,\gamma_2)$ from the initial expressions (g_2 and g_3 are analytical expressions or integrals numerically estimated) in a very dense and appropriately selected grid of (γ_1,γ_2) points; and second, from the $(\gamma_1,\gamma_2,\tau_2)$ and $(\gamma_1,\gamma_2,\tau_3)$ points we form the bivariate linear interpolation functions $\tau_2 = h_2(\gamma_1,\gamma_2)$ and $\tau_3 = h_3(\gamma_1,\gamma_2)$ (note that any mathematical software creates easily bivariate interpolation functions). Replacing τ_2 and τ_3 in these equations with their counterpart estimates $\hat{\tau}_2$ and $\hat{\tau}_3$ we can form a square error norm that can be numerically minimized. Particularly, the estimated shape parameters γ_1 and γ_2 are those emerging by the following expression

$$(\gamma_1, \gamma_2) = \underset{\gamma_1, \gamma_2}{\arg\min} \sum_{j=2}^{3} \left(h_j(\gamma_1, \gamma_2) - \hat{\tau}_j \right)^2 \quad (0)$$

Once the parameters γ_1 and γ_2 are estimated for either distribution the trivial scale parameter β can be directly estimated from the corresponding expression of the first L-moment λ_1 given in Eq. (0) and Eq. (0). As a final technical detail we note that we tested the fitting method to millions of random points to assess its accuracy and to define the parameters' range where the method works essentially without estimation error. As we have observed for the GG distribution these ranges are $0.2 \le \gamma_1 \le 10$ and $0.1 \le \gamma_2 \le 10$, while for the BrXII distribution they are $0.2 \le \gamma_1 \le 10$ and $0.001 \le \gamma_2 \le 0.9$. If the fitting procedure resulted in parameters outside these ranges it was considered inaccurate.

The estimated values of the shape parameters for both distributions are presented in the form of box plots in Figure 11 while some of their basic summary statistics are given in Table 2. Considering the theoretical range of the parameters, i.e., (0, ∞), of both parameters and for both distributions it is apparent that they actually vary in a narrow range as the 95% empirical confidence intervals indicate in Figure 11 (outer fences of the whiskers). For the GG distribution the median of the parameter γ_1 for all months ranges from 1.08 to 1.23 while for all month and for most of the records $\gamma_1 > 1$ indicating bell-shaped densities. The average of all monthly medians of the parameter γ_2 is approximately 0.59 with the majority of records having $\gamma_2 < 1$ indicating a heavier tail than the exponential or the Gamma tail [see also ,21]. The median values of the BrXII γ_1 parameter for all months are close to 1; actually the average of all monthly medians is 0.97, a value very close to the Pareto type II value, i.e., $\gamma_1 = 1$. Additionally, we note that more than 50% of the records have $\gamma_1 < 1$ indicating J-shape densities and verifying also the results presented in Figure 9. Finally, the monthly median values of the γ_2 parameter vary in a narrow range, i.e., form 0.19 to 0.25, while the upper limit in the 95% ECI is for all months (except January) less than 0.5, indicating finite variance distributions.

4.4 Performance of the models

The GG distribution as the analysis showed is able to describe more records than the BrXII. Yet as the two distributions differ significantly in the behavior of the tail, as the former is of exponential form and the latter is power type, it is useful to compare them in terms of some fitting error measures. Obviously, the comparison is possible only for the samples in which both distributions were fitted. For example Figure 12 presents a probability plot of the fitted distributions to the (nonzero) daily precipitation values of a station (station code CA006158350). Clearly, both distributions fit well and it is evident that the BrXII distribution has a heavier tail and thus for small exceedance probabilities (large return periods) predicts larger values.

In order to evaluate and compare the fitting performance of the distributions we define the following four error measures

$$\text{ER-I} = \frac{1}{n}\sum_{i=1}^{n}\left|\Delta x_{(i)}\right| \tag{0}$$

$$\text{ER-II} = \frac{1}{m}\sum_{i=n-m+1}^{n}\left|\Delta x_{(i)}\right| \tag{0}$$

$$\text{ER-III} = \max\left(\left|\Delta x_{(1)}\right|,\ldots,\left|\Delta x_{(n)}\right|\right) \tag{0}$$

$$\text{ER-IV} = \frac{\Delta x_{(n)}}{\hat{x}_{(n)}}100 \tag{0}$$

where $\Delta x_{(i)} = x_{(i)} - \hat{x}_{(i)}$ is the difference between the predicted value $x_{(i)}$ and its corresponding observed one $\hat{x}_{(i)}$ with the index i indicating the position in the ordered sample, i.e., $\hat{x}_{(1)} \le, \ldots, \le \hat{x}_{(n)}$. The predicted value is estimated by the quantile function of each distribution, i.e., $x_{(i)} = Q_X\left(p_i\right)$, using the corresponding empirical probability according to the Weibull plotting position, i.e., $p_i = i/\left(n+1\right)$. Thus, ER-I is the mean value of the absolute differences of all sample values and provides an overall measure of fitting performance; ER-II is focused on the last m largest sample values and may be seen as a fitting measure to the extreme values or to the tail (here we set m = 10); ER-III is the absolute maximum difference identified between observed and predicted values and does not necessarily correspond to the sample's maximum value; ER-IV is focused on the percentage difference between the predicted maximum value and the maximum observed value with negative and positive differences implying, respectively, underestimation or overestimation of the maximum value by the fitted distribution.

The results are presented in Figure 13 (box plots of the four error measures for the values of all months) and in Figure 14 (box plots for the individual months). Additionally, Table 3 shows, for all months and for individual months, the number of records that were actually compared (both distributions fitted) as well as the averages of the error measures. In general, as the box plots and the values of Table 3 reveal, the GG distribution according to all error measures performs better than the BrXII. If we focus on the ER-IV, which estimates the percentage difference between the predicted and the observed maximum value, we note that the GG distribution performs exceptionally well. For example for all months (Figure 13) this estimate is essentially unbiased while the 95% ECI is between −45.6% and 52.2%; in contrast, the BrXII overestimates the maximum on average 28.2% (see Table 3) while the 95% ECI is much wider, i.e., from −35.9% to 120.0%. Yet the performance of the BrXII distribution improves for each specific month separately (Figure 14) where the average overestimation per month for the BrXII is 4.7% (estimated form the values of Table 3) while the GG distribution underestimates on average the maximum value by −2.2%. Finally, the percentage of the records in which the GG distribution was better fitted according to the four error measures are also given in Table 3 while a side-by-side comparison of the two distributions is presented in Figure 15. Apparently, the GG distribution performs better especially according to ER-I which evaluates the overall fitting. Comparing the percentages of the two distributions, shown in Figure 15, we observe that the GG distribution improves even more its performance over the BrXII distribution at the daily precipitation compared to the monthly daily precipitation. This might be an extra argument for the GG distribution as the daily precipitation samples are much larger in size than the monthly samples and thus the parameter estimation is more accurate in this case.

5. Summary and conclusions

In this study we investigate the seasonal variation of daily precipitation focusing on the properties of its marginal distribution. Two were the major questions we tried to answer: (a) which statistical characteristics of daily precipitation vary the most over the months and how much, and (b) whether or not there is a relatively simple probability model that can describe the nonzero daily precipitation at every month and every area of the world. In order to treat these questions we performed a massive analysis of approximately 170,000 monthly daily precipitation records from more than 14,000 stations from all over the globe.

Regarding the first question we first studied the variation of probability dry and of three representative characteristics of the marginal distribution of nonzero daily precipitation, i.e., the mean value, the L-variation and the L-skewness, in the two hemispheres. In general, a typical sinusoidal-like pattern was revealed (see Figure 2) for all statistics and for both hemispheres, with a surprising exception in the probability dry of the SH where a more complicated picture is observed. Additionally, to explore the monthly variation in detail at each record we proposed and applied a test for seasonality, i.e., the SV-Test. Application of the SV-Test revealed a clear monthly variation in probability dry and in the mean value of nonzero daily precipitation in 95.1% and in 91.7%, respectively, of the stations studied (see Figure 5); the corresponding percentages of the shape characteristics, i.e., of L-variation and L-skewness, were 66.1% and 54.2%, respectively, these results if combined with the general picture obtained by the analysis in the hemispheres indicate that, in general, the shape characteristic vary too. The monthly variation of those statistics at each station was quantified by various deviation measures with respect to the average of all months (see Figure 7). The analysis showed that the highest monthly variation is observed in the mean value of nonzero precipitation followed by probability dry, L-skewness and finally by L-variation, implying that although the shape characteristics vary, their variability is much less than of the mean value and the probability dry.

Regarding the second question we tested the performance of two flexible three-parameter distributions: one power-type, the Burr type XII distribution, and one of exponential form, the Generalized Gamma which are generalizations of commonly used two-parameter distributions, e.g., the Pareto, Gamma, Weibull and others. In order to check the suitability of these distributions for the nonzero daily precipitation, first, we used L-moments ratio diagrams to evaluate their potential to describe or reproduce the observed shape characteristics of all records; and second, we actually fitted and estimated the parameters for each distribution and for all records. For the huge number of records analyzed both distributions performed very well. Particularly, the Burr type XII in the worst case, i.e., in November, managed to describe 79.1% of the records (see Figure 9); the corresponding value for the Generalized Gamma distribution was observed in January and was 94.2% while this distribution was able to describe the shape characteristics for all months in 78.8% of the stations (see Figure 10). Finally, the two distributions were compared to each other using various error measures and the Generalized Gamma performed better in most of the cases (see Figure 15).

The implications of this study are: (a) The marginal distribution of daily precipitation varies over the months and over location suggesting the necessity for a flexible probability model. (b) The seasonal and the spatial variability observed in the shape characteristics points out that the commonly used two-parameter models, e.g., the Gamma, the Weibull, the Lognormal, the Pareto, etc. cannot serve as 'universal" models for the daily precipitation. However, we stress that estimating three parameters is more uncertain than estimating two parameters. Thus, if a more parsimonious model is adequate it should always be preferred over a more complicated one. (c) The density function of daily precipitation may significantly differ not only in its general shape, i.e., J-shaped or Bell-shaped, but also in its tail behavior; this dictates that a "universal" probability model for daily precipitation must have in general two shape parameters, one to control the left tail and one to control the right tail. (d) Two simple models with the above characteristics that perform very well are the Burr type XII distribution and the Generalized Gamma distribution with the latter performing even better than the former providing thus an excellent model choice. (e) Using only these two distributions, having some of their characteristics complementary to each other, we can model the entire data set for all months and all stations. (f) The shape parameter γ_2 of the Generalized Gamma distribution, which controls the right tail and thus the extreme values, for the vast majority of records analyzed is $\gamma_2 < 1$, with 1 corresponding to the Gamma distribution; this implies that some of the most commonly used exponential-tail distributions like the Exponential, the Gamma or mixed Exponentials may constitute a dangerous choice and should not be used unjustifiably in practice as they can severely underestimate the magnitude and the frequency of the extreme daily precipitation. (g)

As a rule of thumb, the GG distribution should be the first choice as it is highly likely to provide a good fit to daily precipitation data; if this model is not adequate, the BrXII distribution should be also considered. Finally, given the uncertainty in the estimation of three parameters and the importance of the shape parameter that controls the right tail, in cases where the sample size is small, the mean estimated values could be used a priori, i.e., γ_2 = 0.53 and γ_2 = 0.22 for the GG and the BrXII distributions, respectively. Additionally, a Bayesian method can be used with prior shape parameter distributions based on the statistics provided in Table 2.

Acknowledgment We wish to thank Francesco Laio and an anonymous reviewer for their useful comments which helped in improving the presentation of this study. This research was partially funded by the Greek General Secretariat for Research and Technology through the research project “Combined REnewable Systems for Sustainable Energy DevelOpment” (CRESSENDO, grant number 5145).

References

[1] Smith RL, Schreiber HA. Point Processes of Seasonal Thunderstorm Rainfal 2. Rainfall Depth Probabilities. Water Resources Research 1974.

[2] Todorovic P, Woolhiser DA. A stochastic model of n-day precipitation. Journal of Applied Meteorology 1975;14:17–24.

[3] Woolhiser D, Roldán J. Stochastic daily precipitation models: 2. a comparison of distributions of amounts. Water Resources Research 1982;18:1461–1468.

[4] Wilks DS. Multisite generalization of a daily stochastic precipitation generation model. Journal of Hydrology 1998;210:178–91. doi:10.1016/S0022-1694(98)00186-3.

[5] Wilks DS. Simultaneous stochastic simulation of daily precipitation, temperature and solar radiation at multiple sites in complex terrain. Agricultural and Forest Meteorology 1999;96:85–101. doi:10.1016/S0168-1923(99)00037-4.

[6] Buishand TA. Some remarks on the use of daily rainfall models. Journal of Hydrology 1978;36:295–308. doi:10.1016/0022-1694(78)90150-6.

[7] Bruhn JA, Fry WE, Fick GW. Simulation of Daily Weather Data Using Theoretical Probability Distributions. Journal of Applied Meteorology 1980;19:1029–36. doi:10.1175/1520-0450(1980)019<1029:SODWDU>2.0.CO;2.

[8] Geng S, Penning de Vries FWT, Supit I. A simple method for generating daily rainfall data. Agricultural and Forest Meteorology 1986;36:363–76. doi:10.1016/0168-1923(86)90014-6.

[9] Swift LW, Schreuder HT. Fitting Daily Precipitation Amounts Using the SB Distribution. Monthly Weather Review 1981;109:2535–40. doi:10.1175/1520-0493(1981)109<2535:FDPAUT>2.0.CO;2.

[10] Wilson PS, Toumi R. A fundamental probability distribution for heavy rainfall. Geophysical Research Letters 2005;32:L14812.

[11] Biondini R. Cloud Motion and Rainfall Statistics. Journal of Applied Meteorology 1976;15:205–24. doi:10.1175/1520-0450(1976)015<0205:CMARS>2.0.CO;2.

[12] Shimizu K. A bivariate mixed lognormal distribution with an analysis of rainfall data. Journal of Applied Meteorology;(United States) 1993;32.

[13] Mielke Jr PW. Another Family of Distributions for Describing and Analyzing Precipitation Data. Journal of Applied Meteorology 1973;12:275–80.

[14] Mielke Jr PW, Johnson ES. Three-Parameter Kappa Distribution Maximum Likelihood Estimates and Likelihood Ratio Tests. Monthly Weather Review 1973;101:701–7.

[15] Hosking JRM. The four-parameter kappa distribution. IBM Journal of Research and Development 1994;38:251–258.

[16] Park J-S, Seo S-C, Kim TY. A kappa distribution with a hydrological application. Stoch Environ Res Risk Assess 2009;23:579–86. doi:10.1007/s00477-008-0243-5.

[17] Mielke Jr PW, Johnson ES. Some generalized beta distributions of the second kind having desirable application features in hydrology and meteorology. Water Resources Research 1974;10:223–226.

[18] Fitzgerald DL. Single station and regional analysis of daily rainfall extremes. Stochastic Hydrol Hydraul 1989;3:281–92. doi:10.1007/BF01543461.

[19] Papalexiou SM, Koutsoyiannis D. Entropy based derivation of probability distributions: A case study to daily rainfall. Advances in Water Resources 2012;45:51–7. doi:10.1016/j.advwatres.2011.11.007.

[20] Papalexiou SM, Koutsoyiannis D. The battle of extreme value distributions: A global survey on the extreme daily rainfall. Water Resources Research 2012. doi:10.1029/2012WR012557.

[21] Papalexiou SM, Koutsoyiannis D, Makropoulos C. How extreme is extreme? An assessment of daily rainfall distribution tails. Hydrology and Earth System Sciences Discussions 2012;9:5757–78. doi:10.5194/hessd-9-5757-2012.

[22] Hosking JRM. L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics. Journal of the Royal Statistical Society Series B (Methodological) 1990;52:105–24.

[23] Hosking JRM. Moments or L Moments? An Example Comparing Two Measures of Distributional Shape. The American Statistician 1992;46:186–9. doi:10.2307/2685210.

[24] Kottek M, Grieser J, Beck C, Rudolf B, Rubel F. World Map of the Koppen-Geiger climate classification updated. Meteorologische Zeitschrift 2006;15:259–63. doi:10.1127/0941-2948/2006/0130.

[25] Peel MC, Finlayson BL, McMahon TA. Updated world map of the Köppen-Geiger climate classification. Hydrol Earth Syst Sci 2007;11:1633–44. doi:10.5194/hess-11-1633-2007.

[26] Burr IW. Cumulative Frequency Functions. The Annals of Mathematical Statistics 1942;13:215–32.

[27] Tadikamalla PR. A Look at the Burr and Related Distributions. International Statistical Review / Revue Internationale de Statistique 1980;48:337–44.
[28] Stacy EW. A Generalization of the Gamma Distribution. The Annals of Mathematical Statistics 1962;33:1187–92.
[29] Vogel RM, Fennessey NM. L moment diagrams should replace product moment diagrams. Water Resources Research 1993;29:1745–1752.
[30] Peel MC, Wang Q, Vogel RM, McMahon TA. The utility of L-moment ratio diagrams for selecting a regional probability distribution. Hydrological Sciences Journal 2001;46:147–156.

Tables

Table 1. Mean values and standard deviation values of the four statistics studied.

		All	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Northern hemisphere														
P_{dry}	mean	72.03	73.55	74.23	74.03	73.18	71.05	68.49	67.80	68.97	71.37	74.70	73.68	73.65
	SD	11.19	16.74	15.10	14.24	13.28	12.71	13.48	15.95	15.30	12.78	13.50	16.45	17.36
μ	mean	9.52	7.08	7.18	7.80	8.28	8.99	9.95	10.21	10.11	10.47	10.04	8.73	7.58
	SD	4.67	4.31	4.26	4.25	4.14	4.31	4.86	5.22	4.70	4.94	5.20	4.93	4.57
τ_2	mean	0.59	0.56	0.56	0.56	0.57	0.57	0.58	0.58	0.59	0.59	0.59	0.57	0.57
	SD	0.04	0.05	0.05	0.05	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.05	0.05
τ_3	mean	0.46	0.44	0.43	0.43	0.43	0.43	0.44	0.45	0.46	0.46	0.45	0.44	0.44
	SD	0.05	0.06	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.06
Southern hemisphere														
P_{dry}	mean	77.91	77.73	76.80	78.29	79.69	78.32	76.50	76.84	77.91	78.77	77.74	77.85	78.05
	SD	10.60	14.38	14.34	12.96	11.62	13.35	15.99	17.32	16.79	14.25	12.22	12.06	13.37
μ	mean	9.27	11.09	11.46	10.54	9.06	8.34	7.71	7.21	6.81	7.15	8.21	9.01	10.08
	SD	3.70	4.56	4.47	4.22	3.55	3.22	3.19	2.98	2.62	2.74	3.15	3.53	4.04
τ_2	mean	0.58	0.59	0.59	0.59	0.58	0.58	0.58	0.57	0.56	0.56	0.56	0.56	0.57
	SD	0.05	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.05	0.05	0.05	0.05
τ_3	mean	0.46	0.47	0.47	0.47	0.46	0.46	0.45	0.45	0.44	0.44	0.44	0.44	0.45
	SD	0.06	0.07	0.07	0.07	0.07	0.07	0.08	0.08	0.08	0.07	0.07	0.07	0.07
Global														
P_{dry}	mean	74.40	75.24	75.27	75.75	75.80	73.99	71.72	71.44	72.58	74.36	75.92	75.36	75.42
	SD	11.33	15.97	14.85	13.90	13.04	13.45	15.06	17.10	16.51	13.87	13.08	14.98	16.01
μ	mean	9.42	8.70	8.91	8.90	8.60	8.73	9.05	9.00	8.78	9.13	9.30	8.85	8.59
	SD	4.31	4.83	4.83	4.45	3.93	3.92	4.41	4.69	4.31	4.50	4.57	4.42	4.53
τ_2	mean	0.58	0.57	0.57	0.57	0.57	0.57	0.58	0.58	0.58	0.58	0.57	0.57	0.57
	SD	0.04	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
τ_3	mean	0.46	0.45	0.45	0.45	0.44	0.44	0.45	0.45	0.45	0.45	0.44	0.44	0.44
	SD	0.05	0.07	0.07	0.07	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.07

Table 2. Basic summary statistics of the estimated shape parameters of the GG and BrXII distributions.

	All	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
GG distribution													
Fit No.	13826	12729	13012	13116	13353	13445	13491	13292	13317	13509	13620	13410	13000
Parameter γ_1													
Q_{50}	1.20	1.23	1.22	1.17	1.13	1.09	1.08	1.09	1.10	1.09	1.10	1.13	1.21
μ	1.50	1.63	1.59	1.53	1.45	1.39	1.36	1.41	1.43	1.41	1.42	1.49	1.61
σ	0.94	1.22	1.15	1.07	1.00	0.97	0.94	1.01	1.04	1.02	1.02	1.11	1.20
τ_2	0.29	0.34	0.33	0.32	0.31	0.30	0.30	0.31	0.32	0.31	0.31	0.33	0.34
τ_3	0.38	0.43	0.42	0.42	0.42	0.43	0.43	0.43	0.44	0.44	0.43	0.43	0.42
Parameter γ_2													
Q_{50}	0.52	0.54	0.54	0.58	0.61	0.62	0.61	0.60	0.59	0.59	0.60	0.60	0.56
μ	0.53	0.58	0.58	0.59	0.62	0.62	0.62	0.61	0.60	0.60	0.61	0.63	0.60
σ	0.22	0.30	0.31	0.28	0.28	0.26	0.27	0.28	0.27	0.27	0.28	0.32	0.31
τ_2	0.23	0.28	0.28	0.26	0.25	0.23	0.23	0.24	0.24	0.23	0.24	0.26	0.28
τ_3	0.06	0.14	0.14	0.08	0.06	0.04	0.08	0.09	0.09	0.10	0.10	0.12	0.13
Butt XII distribution													
Fit No.	12744	11900	11827	11810	11555	11460	11544	11737	11878	11768	11503	11203	11551
Parameter γ_1													
Q_{50}	0.94	1.00	0.98	0.98	0.97	0.96	0.95	0.95	0.95	0.95	0.96	0.99	1.01
μ	0.96	1.05	1.03	1.01	1.00	0.99	0.98	0.99	0.99	0.98	0.99	1.02	1.05
σ	0.16	0.24	0.23	0.21	0.18	0.18	0.19	0.21	0.20	0.19	0.19	0.23	0.24
τ_2	0.09	0.12	0.12	0.11	0.10	0.10	0.10	0.11	0.11	0.10	0.10	0.11	0.11
τ_3	0.14	0.21	0.21	0.20	0.18	0.19	0.21	0.22	0.19	0.19	0.16	0.18	0.19
Parameter γ_2													
Q_{50}	0.21	0.25	0.24	0.22	0.20	0.19	0.19	0.20	0.20	0.19	0.20	0.21	0.24
μ	0.22	0.25	0.24	0.23	0.22	0.21	0.20	0.21	0.21	0.21	0.21	0.22	0.24
σ	0.11	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.12	0.12	0.13	0.13
τ_2	0.30	0.30	0.30	0.33	0.35	0.36	0.35	0.35	0.34	0.33	0.33	0.32	0.31
τ_3	0.02	0.05	0.04	0.07	0.09	0.11	0.12	0.12	0.12	0.10	0.09	0.07	0.04

Table 3. Mean values of the error measures evaluating the fitting performance of the distributions, as well as percentage values of records in which the GG was better fitted compared to Burr XII.

	All	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Fit No.	12413	10474	10684	10769	10754	10750	10879	10877	11041	11124	10967	10457	10396
Mean values of the error measures for the GG distribution													
ER-I	1.4	1.0	1.0	1.0	0.9	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0
ER-II	14.1	5.5	5.5	5.5	4.9	5.2	5.7	5.9	5.8	5.9	5.5	5.0	5.1
ER-III	38.2	18.9	18.6	19.0	17.0	17.8	20.2	20.1	20.0	20.3	19.5	17.5	17.8
ER-IV	0.7	-1.6	-1.6	-2.2	-2.1	-1.7	-2.7	-1.7	-2.4	-2.7	-3.1	-2.2	-2.2
Mean values of the error measures for Burr XII distribution													
ER-I	2.2	1.1	1.2	1.1	1.0	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1
ER-II	25.4	5.8	5.9	5.9	5.2	5.6	6.1	6.3	6.2	6.1	5.8	5.3	5.4
ER-III	62.0	19.8	19.9	20.1	17.9	18.8	21.0	21.3	20.9	20.9	20.1	18.2	18.6
ER-IV	28.2	5.8	5.2	4.5	4.2	4.9	4.2	5.5	4.7	4.1	3.6	4.6	5.0
Percentage the GG distribution better fitted compared to Burr XII (%)													
ER-I	87.0	80.8	80.9	77.9	77.8	76.2	74.6	79.6	77.6	75.4	77.1	78.0	78.4
ER-II	79.2	65.8	66.1	62.9	62.5	63.3	61.3	65.2	63.2	59.3	60.6	63.6	64.9
ER-III	69.5	59.9	60.2	56.6	56.4	56.8	55.1	58.7	56.8	54.1	54.1	58.3	58.6
ER-IV	67.0	55.8	55.8	53.1	53.9	53.3	52.5	55.5	54.2	52.5	51.7	54.9	55.3

Figures

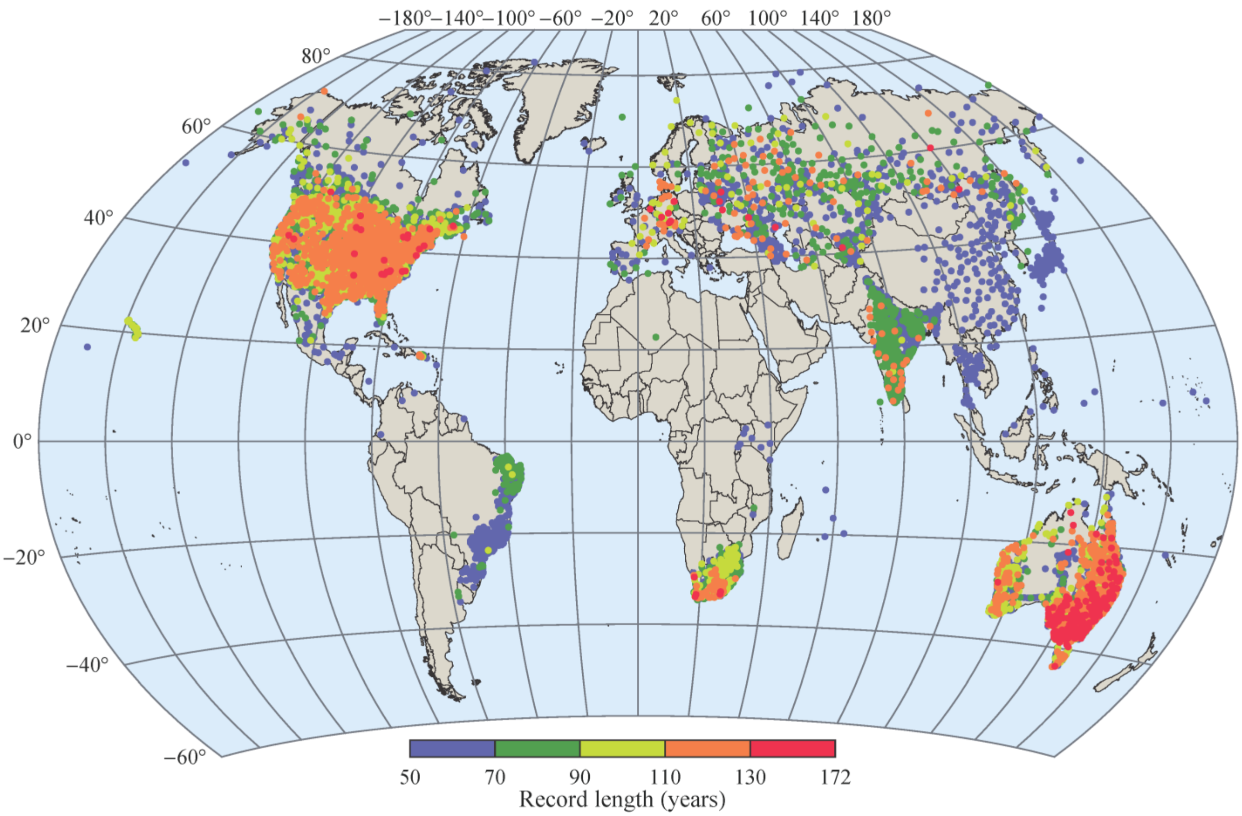

Figure 1. Locations of the 14,157 stations studied.

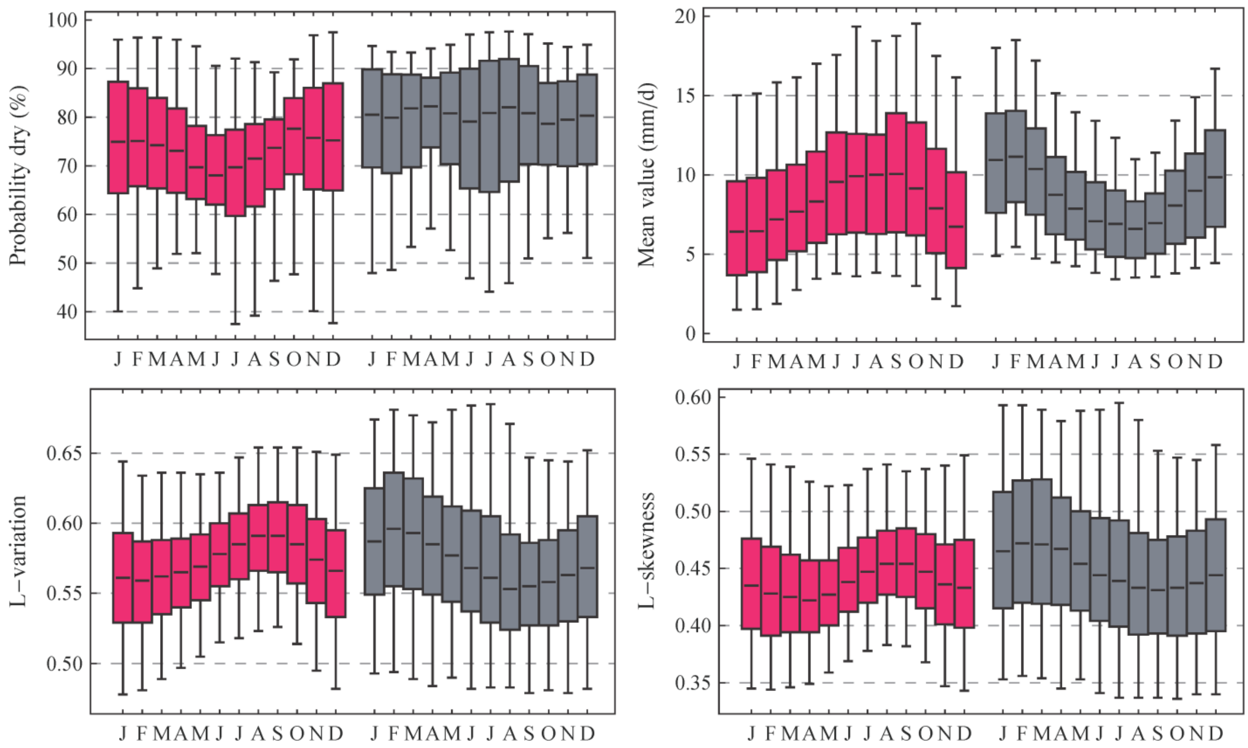

Figure 2. Estimated statistics of the monthly daily records analyzed; red box plots on the left are for the NH; gray boxplots on the right are for the SH; outer fences indicate the 90% ECI.

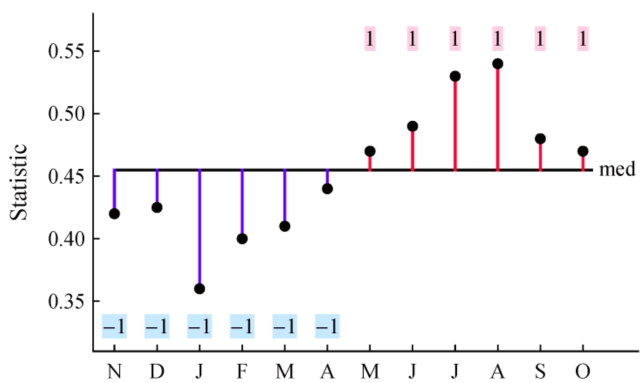

Figure 3. Explanatory sketch of the seasonal variation test; values above and below the median are denoted, respectively, with 1 and −1.

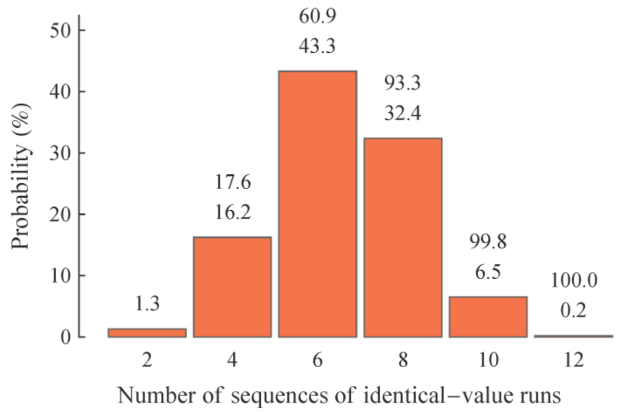

Figure 4. Benchmark values for the SV-Test; the bars indicate the probabilities (the upper number is cumulative) corresponding to specific number of SIVR in the case of 12 randomly generated numbers (no seasonality).

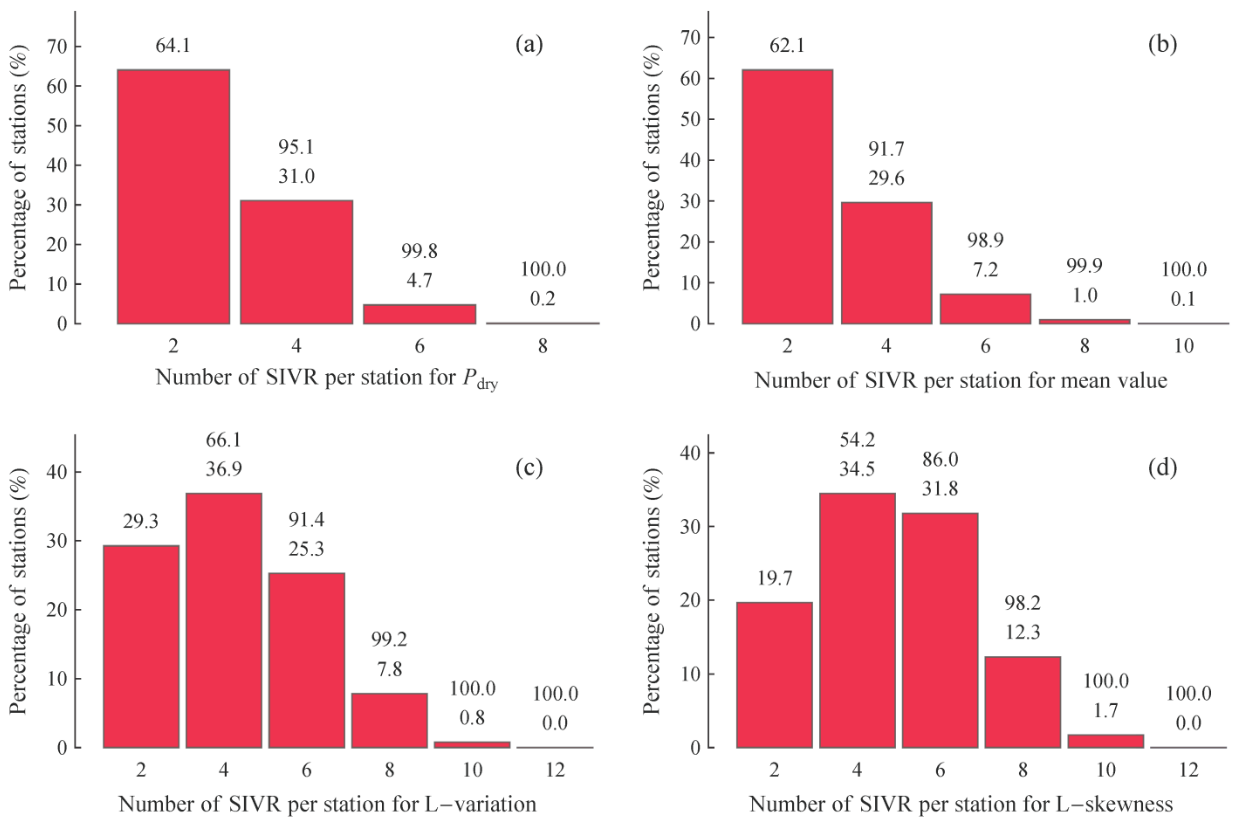

Figure 5. Results of the SV-Test applied to: (a) the probability dry, (b) mean value (c) L-variation and (d) L-skewness.

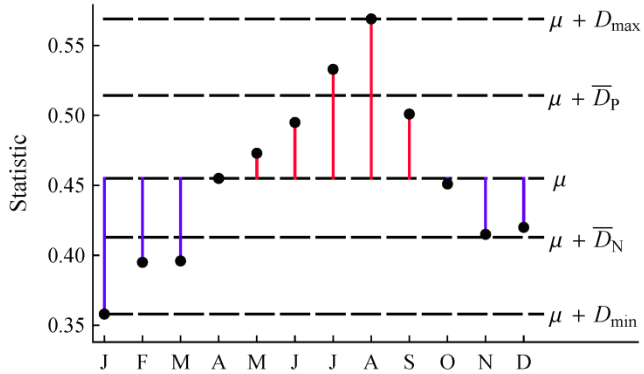

Figure 6. Explanatory sketch of the four difference measures studied.

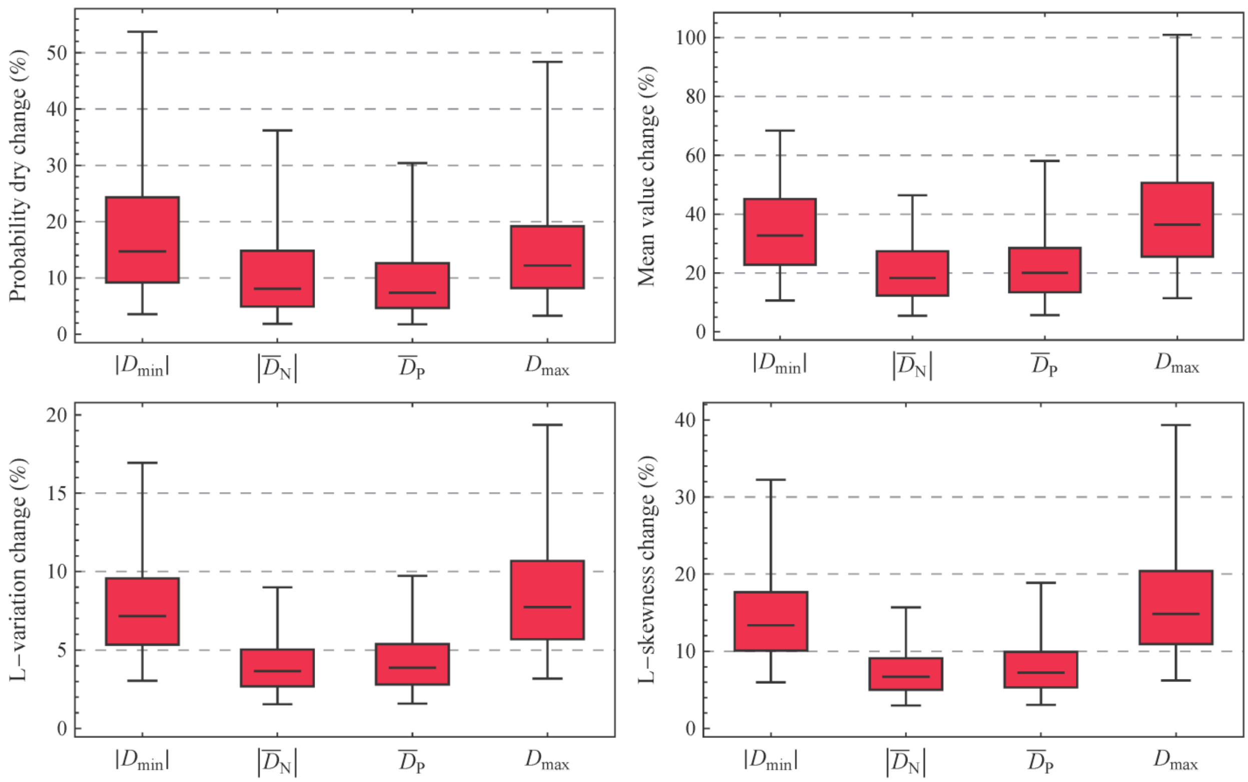

Figure 7. Box plots depicting the percentage change of the difference measures relative to the average of all months for the four statistics studied. Each box plot is constructed by the values determined from the stations studied. Outer fences indicate the 95% ECI.

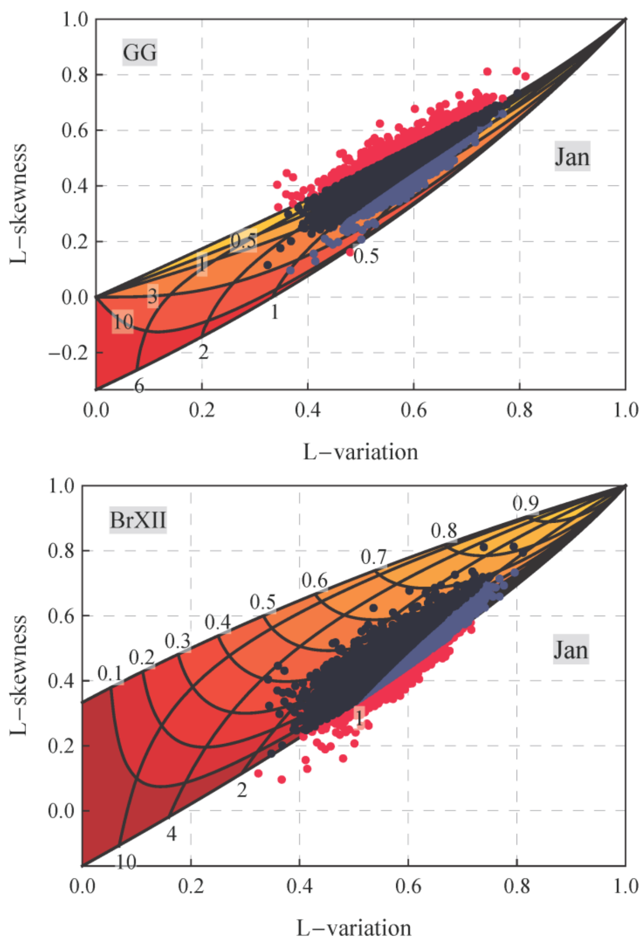

Figure 8. Observed L-points for the month of January of the 14,157 daily precipitation records studied in comparison to the theoretical L-areas of (a) the BrXII distribution and (b) the GG distribution. Red-colored L-points lie outside the L-area; dark-colored indicate a Bell-shaped distribution; light-colored indicate a J-shaped distribution.

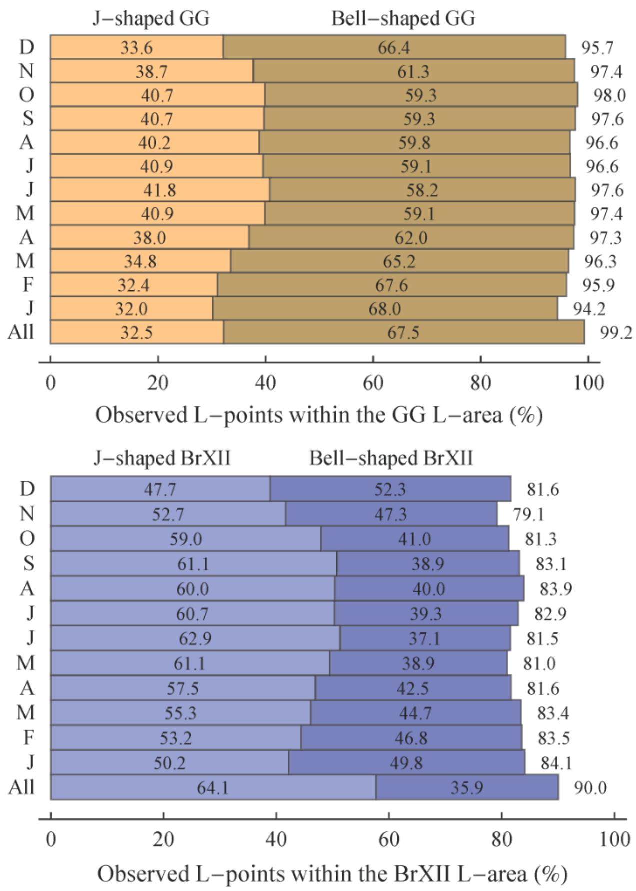

Figure 9. Percentage of empirical L-points lying within the L-areas of the GG and the BrXII distributions.

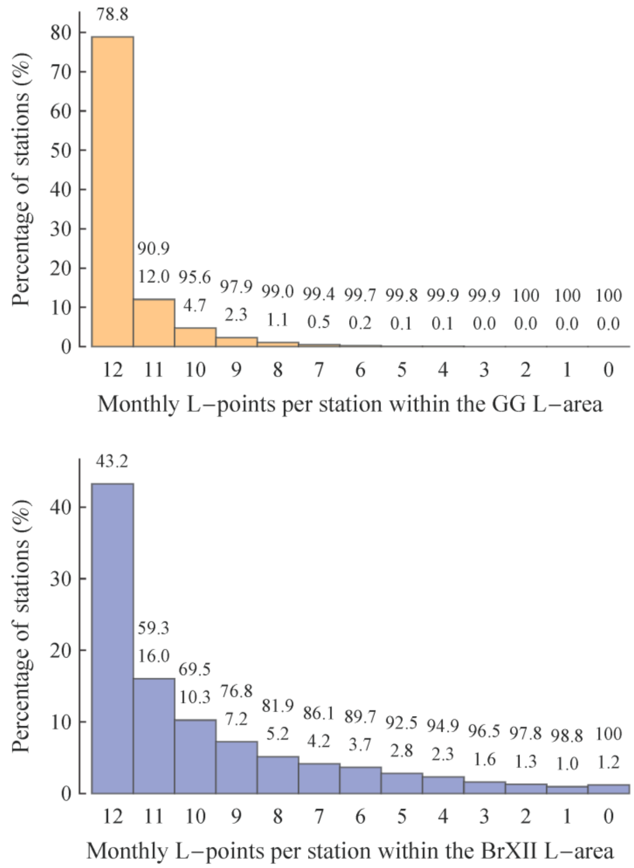

Figure 10. Percentage of records *vs*. the number of monthly L-points per station lying within the theoretical L-areas of the GG and the BrXII distributions.

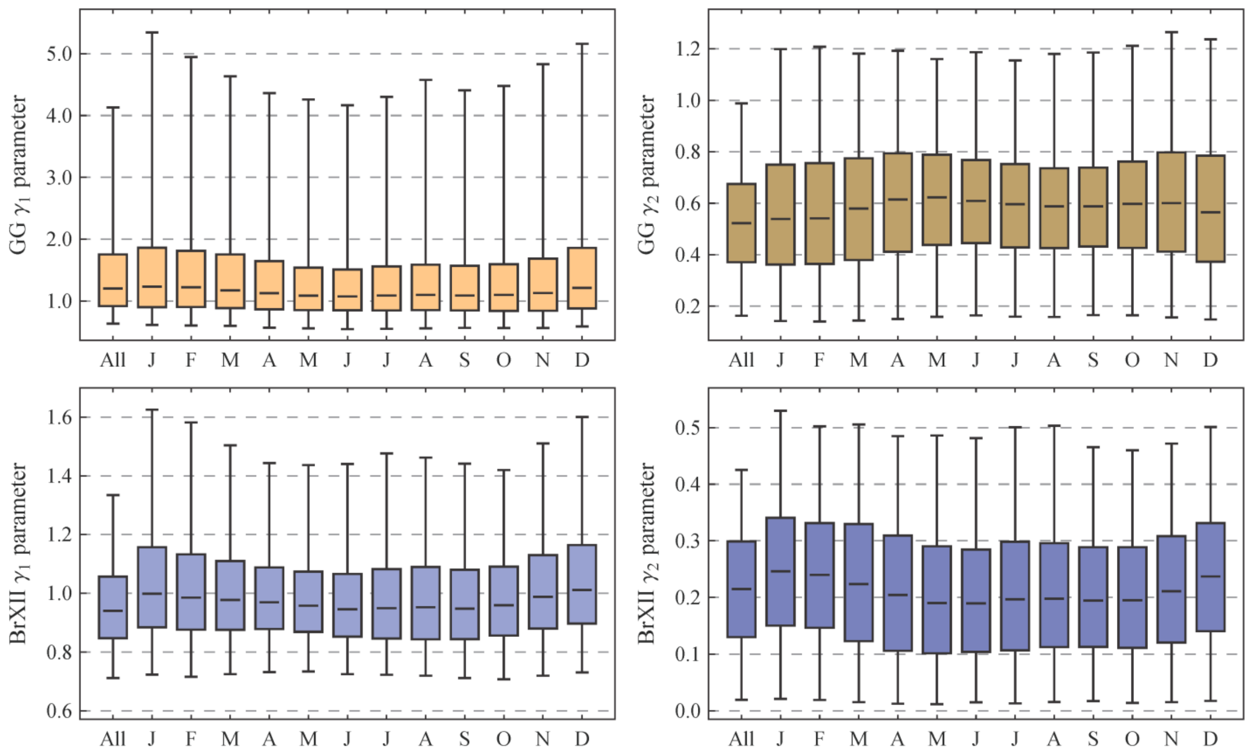

Figure 11. Estimated shape parameters of the GG and BrXII distributions using the method of L-moments.

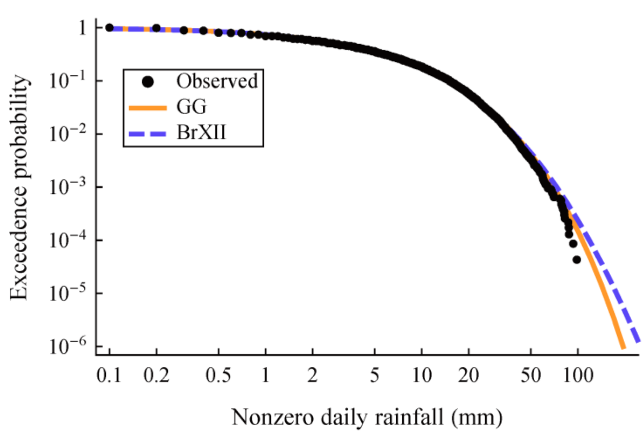

Figure 12. Probability plot of the fitted distributions to a specific station (station code CA006158350) using the method of L-moments.

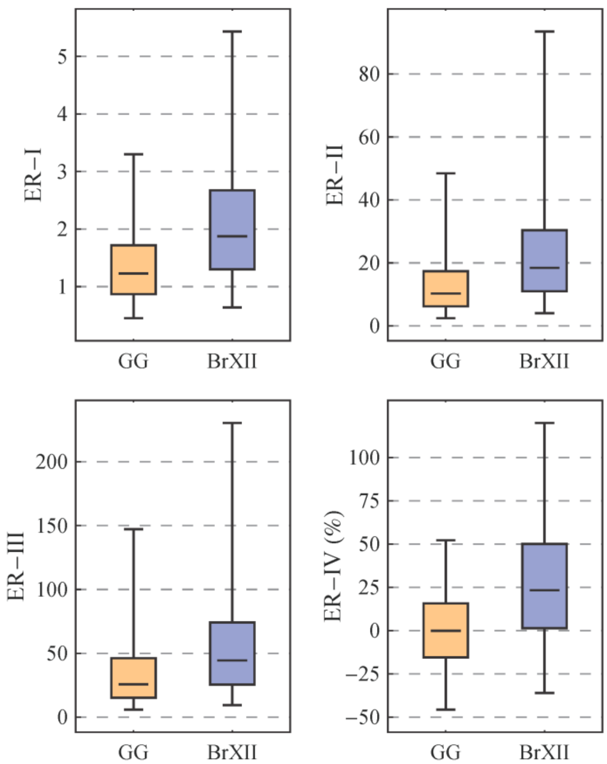

Figure 13. Box plots of the error measures that evaluate the fitting performance of the GG and BrXII distributions to daily precipitation of all months.

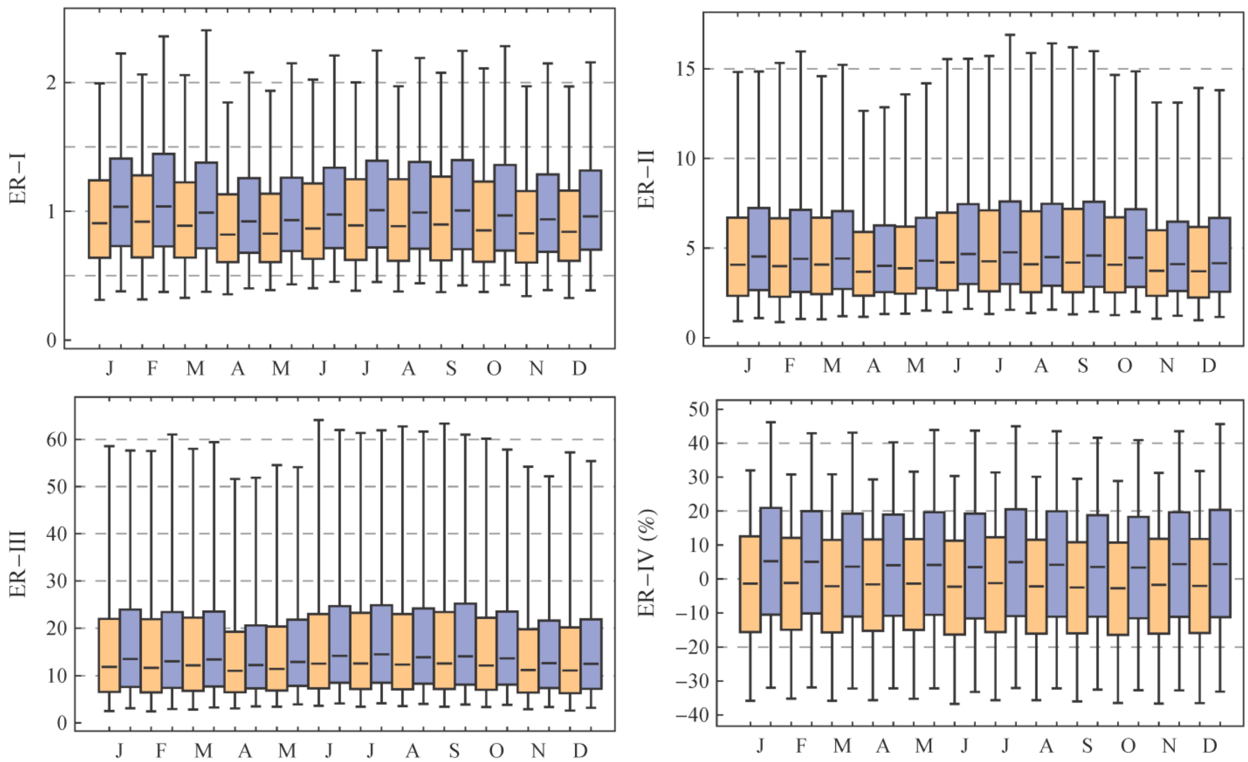

Figure 14. Box plots of the error measures of the fitting of the GG and BrXII distributions to the monthly daily precipitation records.

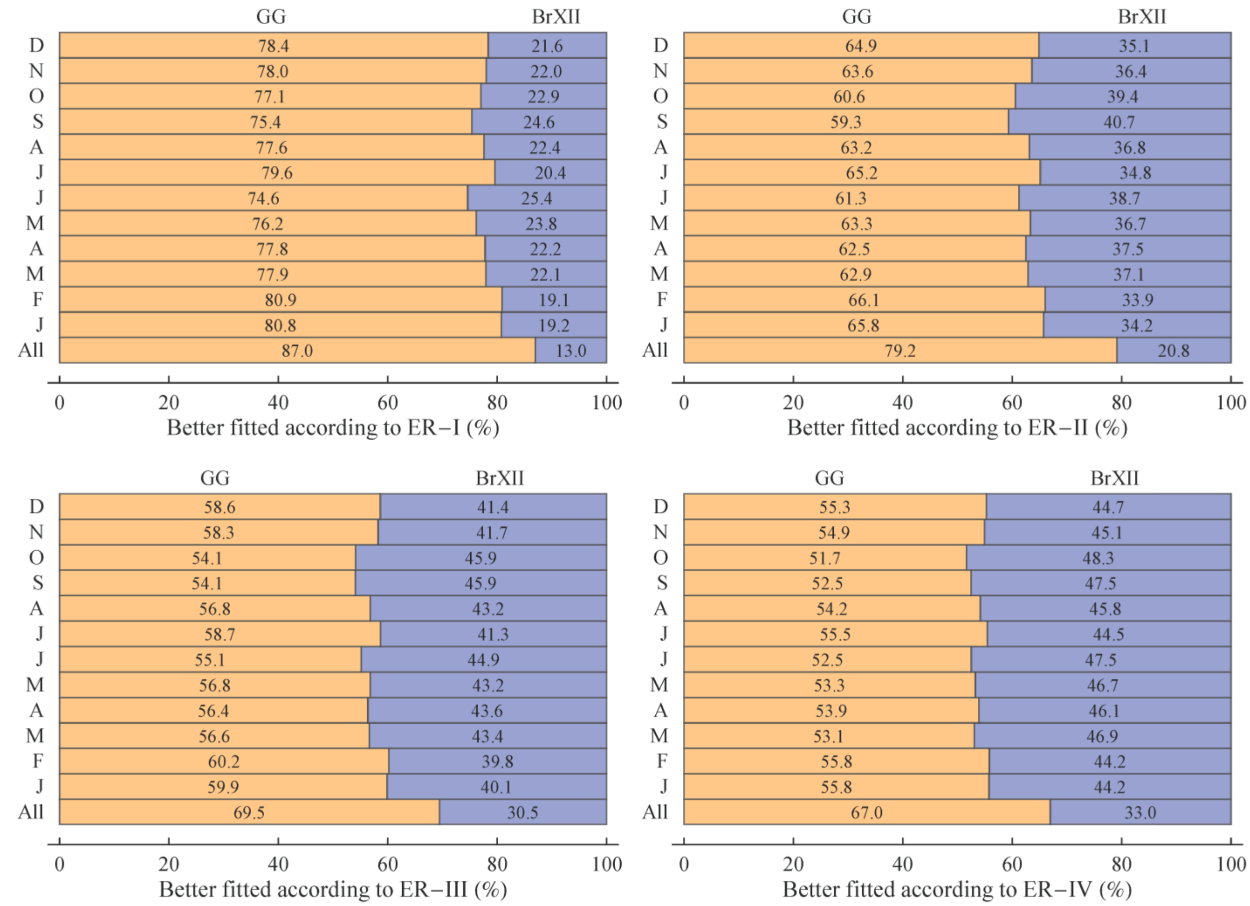

Figure 15. Comparison of the fitting performance of the two distributions; the values within the bars indicate the percentage of stations in which each distribution was better fitted according to the error measures.