AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2014WR015874

**Key Points:**
- Compute entropy for one-dimensional case
- Four issues of practical relevance are considered
- Accuracy and robustness are tested

**Correspondence to:**
W. Gong,
gongwei2012@bnu.edu.cn

**Citation:**
Gong, W., D. Yang, H. V. Gupta, and G. Nearing (2014), Estimating information entropy for hydrological data: One-dimensional case, *Water Resour. Res.*, *50*, 5003–5018, doi:10.1002/2014WR015874.

Received 19 MAY 2014
Accepted 25 MAY 2014
Accepted article online 6 JUN 2014
Published online 19 JUN 2014

# Estimating information entropy for hydrological data: One-dimensional case

**Wei Gong[1,2], Dawen Yang[2], Hoshin V. Gupta[3], and Grey Nearing[4]**

[1]College of Global Change and Earth System Science, Beijing Normal University, Beijing, China, [2]State Key Laboratory of Hydroscience and Engineering, Tsinghua University, Beijing, China, [3]Department of Hydrology and Water Resources, University of Arizona, Tucson, Arizona, USA, [4]Science Applications International Corporation, NASA Goddard Space Flight Center, Hydrologic Sciences Lab, Greenbelt, Maryland, USA

**Abstract** There has been a recent resurgence of interest in the application of *Information Theory* to problems of system identification in the *Earth and Environmental Sciences*. While the concept of entropy has found increased application, little attention has yet been given to the practical problems of estimating entropy when dealing with the unique characteristics of two commonly used kinds of hydrologic data: rainfall and runoff. In this paper, we discuss four important issues of practical relevance that can bias the computation of entropy if not properly handled. The first (*zero effect*) arises when precipitation and ephemeral streamflow data must be viewed as arising from a discrete-continuous hybrid distribution due to the occurrence of many zero values (e.g., days with no rain/no runoff). Second, in the widely used bin-counting method for estimation of PDF's, significant error can be introduced if the bin width is not carefully selected. The third (*measurement effect*) arises due to the fact that continuously varying hydrologic variables can typically only be observed discretely to some degree of precision. The Fourth (*skewness effect*) arises when the distribution of a variable is significantly skewed. Here we present an approach that can deal with all four of these issues, and test them with artificially generated and real hydrological data. The results indicate that the method is accurate and robust.

## 1. Introduction

The concept of entropy has found widespread application to various problems in hydrology and water resources [*Singh*, 1997, 2000, 2013]. While the notion of "*entropy*" originates in physics, Shannon used the term to define a metric that characterizes and quantifies the degree of uncertainty (or lack of information) about some quantity [*Shannon*, 1948]. In the recent hydrology and water resources literature, there is a resurgence of interest in the application of entropy. *Gong et al*. [2013] have shown that a process based on *Independent Component Analysis* (ICA) [*Hyvarinen and Oja*, 2000], the "*Independence Rule*" and the "*Chain Rule*" (Theorem 8.6.2 in *Cover and Thomas* [2006]), can be used to bring methods of 1-D entropy estimation to bear on the more challenging problems of estimating high-dimensional entropy, mutual information, and other metrics such as transfer entropy and KL-divergence. With this technology, we can push forward many existing investigations, such as ecohydrological processes network analysis based on transfer entropy [*Ruddell and Kumar*, 2009a, 2009b], input variable selection for seasonal and interannual rainfall probabilistic forecasts [*Sharma et al*., 2000; *Sharma*, 2000a, 2000b; *Fernando et al*., 2009; *May et al*., 2008], and quantification of the skill of ensemble forecasts [*Weijs et al*., 2010a, 2010b; *Weijs and van de Giesen*, 2011].

The strategy used by most hydrologists to compute entropy is intuitive: first estimate the PDF (probability density function) of the data, then compute entropy from the definition. A number of different methods for estimating the PDF exists, including: (1) construction of the frequency histograms (also known as *Bin Counting*), (2) the *Naive* estimator, (3) the *Kernel Density* estimator, (4) the *Nearest Neighbor* method, (5) the *Variable Kernel* method, (6) *Orthogonal Series* estimators, (7) *Maximum Penalized Likelihood* estimators, (8) *General Weight Function* estimators, and so on; a nice summary can be found in *Silverman* [1986]. Of these, the three most widely used are the *Bin-Counting, Kernel Density*, and *Average Shifted Histogram* (ASH) estimators [*Scott*, 2008].

While new technologies are constantly being developed, we are currently faced with some practical problems associated with rainfall and runoff data due to their special characteristics: (1) due to the occurrence of

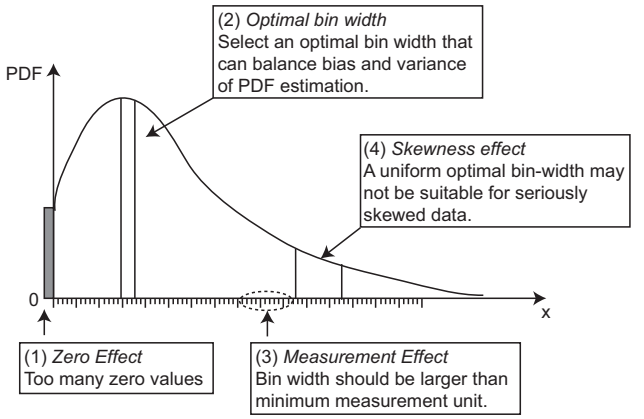


**Figure 1.** A conceptual figure about the four problems going to be discussed in this paper.

many zero values, precipitation and ephemeral streamflow data must be viewed as a discrete-continuous hybrid distribution; (2) the bin-width, or more generally, smoothing parameters, should be carefully selected; (3) the minimum measurement unit should be considered when selecting bin-width; (4) the skewness of rainfall and runoff data may introduce biasedness to PDF and entropy estimation. It is quite easy to demonstrate that inadequate attention to these problems can lead to nonnegligible errors in density and entropy estimation, as shown in the intercomparison in section 4. In this paper, we want to warn about some oft-overlooked problems of PDF and entropy estimation. In the past, scientists have paid attention to these problems including (for example) Chapman's work dealing with zero values in precipitation data [*Chapman*, 1986] and Scott's work about optimal bin width [*Scott*, 1979]. However, these have not received much attention recently, with the exception of *Pechlivanidis et al*. [2012a, 2012b] and I. G. Pechlivanidis et al. (Use of an entropy-based metric in multi-objective calibration to improve model performance, submitted to *Water Resources Research*, 2013, Robust informational entropy-based descriptors of flow in catchment hydrology, submitted to *Hydrologic Sciences Journal*, 2013) who recognize and discuss problems associated with proper selection of bin-width, skewness, and measurement truncation error associated with streamflow. In this paper, we address in details the four typical problems in PDF and entropy estimation for rainfall and runoff data (see Figure 1), and present practical methods for dealing with them.

The rest of this paper is organized as follows: section 2 discusses some special characteristics of rainfall and runoff data that must be accounted for in the computation of *Information Theory* metrics. Section 3 develops a practical methodology for computing accurate estimates of 1-D random variable's entropy; we do not address higher-dimensional PDF's in this paper. Section 4 presents two case studies—one with synthetic and the other with real data—illustrating the issues discussed herein. Finally, section 5 discusses the significance of this work to ongoing research and suggests some possible applications to hydrological investigation.

## 2. Four Problems of PDF and Entropy Estimation

### 2.1. Zero Effect

Precipitation records usually contain a large number of zero values corresponding to periods of no rain. For example, in the widely used Leaf River daily rainfall data set [*Sorooshian et al*., 1993], the fraction of days without rainfall in a 40 water-year period (October 1948 to September 1988) is 54.89%. Given that the Leaf River basin is humid, being located in the state of Mississippi, basins in arid and semiarid areas will have correspondingly more days without rainfall. Similarly, discharge records for ephemeral rivers also have many zero values. Because of the large fraction of zero values in such data, it becomes necessary to handle the nonzero and zero values separately in the process of entropy estimation. *Chapman* [1986] presented some pioneering work that uses a delta function to deal with the zero effect; his approach is clean and robust but is, unfortunately, not very well known. Here we use this approach to deal with the "zero effect."

### 2.2. Optimal Bin Width

Among the three commonly used methods for estimating the form of the data PDF, the *Bin-Counting* and *ASH* methods both require selection of a "*bin width*," while the *Kernel Density* method requires specification of a "*smoothing parameter*." In fact, as discussed by *Scott* [2008], all nonparametric methods for PDF estimation can be asymptotically viewed as "Kernel" methods, and the bin width can be understood as a special

kind of smoothing parameter—one that treats the data points falling in each bin as being from a locally uniform distribution.

For the simple *Bin-Counting* method, the selection of bin width is quite important. Too small a bin width may lead to a histogram that is too rough an approximation of the underlying distribution, while an overly large bin width may result in a histogram that is overly smooth compared to the true pdf. If the histogram is too "rough," the variance of estimated PDF will be too high, whereas if the histogram is overly smooth this will introduce a significant amount of bias. As pointed out by *Scott* [1979] and *Scott and Thompson* [1983], it is necessary to find a compromise value for the optimal bin width that balances variance and bias errors. Unfortunately, Scott's work is also not well known among hydrologists.

Given that *Bin Counting* is commonly used in hydrology, it seems necessary to clarify the importance of selecting the optimal bin width and how that should be achieved. Similarly, when using the *ASH* or *Kernel Density* method, the smoothing parameters must be optimally selected. A comprehensive study of optimal bin width and smoothing parameter selection appears in chapters 3, 5, and 6 of *Scott* [2008]; in section 3, we present an outline of their main conclusions as a quick reference for the reader.

### 2.3. Measurement Effect

Generally speaking, precipitation and river discharge data are usually NOT represented in continuous fashion due to the fact that measurement devices typically record the values in a discrete way. For example, precipitation gauges are typically calibrated in discrete increments, which might be done using a standard cylinder, or using a measuring rod graduated at 0.1 mm intervals. In this case, precipitation depths between 0.05 and 0.1 mm may be rounded up to 0.1 mm and values smaller than 0.05 mm recorded as "*trace*" amounts [see *Anderson and Mcdonnell*, 2005, section 35]. The fact that the data for these "*continuously varying*" quantities are presented discretely with finite precision needs to be taken into account when estimating entropy metrics, as will be explained later. Further, setting the bin width to be equal to the precision of the observations may not provide the best estimate. As discussed in section 2.2, the bin width should be neither too small nor too large. Being subjected to the optimal bin width, the precision of the observation can serve as a guide (or limit) to the selection of the bin width.

Further, the minimum measurement unit only represents one aspect of observation error. Various observation errors, such as rainfall measurement error caused by spatial heterogeneity and streamflow measurement error cause by the uncertainty of stage-discharge relationship, are heteroscedastic, that is to say the expected size of the error tends to increase with the magnitude of rainfall/runoff [*Sorooshian and Dracup*, 1980]. Heteroscedastic error has long been discussed for many years and there are many method to handle it, such as the likelihood function proposed by *Schoups and Vrugt* [2010] and Box-Cox transformation referred in *Misirli et al.* [2003]. In this paper, we use the Box-Cox transformation to reduce the influence of heteroscedasticity.

### 2.4. Skewness Effect

Another characteristic feature of rainfall and runoff is the preponderance of many smaller values and the relative scarcity of larger values (due to their smaller frequency of occurrence). This introduces the problem that a single "*optimal*" bin width may be inappropriate when computing the entropy of the data set. The optimal bin-width derived and proposed by *Scott* [1979] is for a Gaussian distribution, although the paper also discusses the influence of skewness and kurtosis for the cases of lognormal and Student's t distributions. To deal with skewness and measurement truncation of river discharge data, Pechlivanidis et al. (submitted manuscript, 2013) introduced a novel binning method to compute an importance-weighted entropy-based model performance measure. To theoretically determine the influence of skewness on the optimal choice of bin width, one must assume an analytical form for the data PDF but, in practice, this will not generally be known. We propose later to handle this problem in a simple manner by transforming the raw data to a distributional form that is "*approximately normal*," which would then allow us to use the established theory regarding optimal bin widths for Gaussian data. In this case, we need to account for the effects of the distributional transformation; i.e., we need to know the relationship between the entropy values before and after transformation.

## 3. Methodology

We develop an entropy estimator that integrates the following features for rainfall and runoff data: (1) Chapman's Dirac delta function to deal with zero values; (2) Scott's works about optimal bin width; (3)

consider the minimum measurement unit and make sure the optimal bin width is not less than that; (4) apply Box-Cox transformation to reduce the influence of skewness; and (5) use bootstrap method to give an estimation of confidence interval and investigate the influence of small sample size.

### 3.1. Discrete and Continuous Entropy

The definitions (and computation) of entropy are different for the cases of discrete and continuous random variables. For a random variable $X_d$ that can take on only $N_d$ discrete values $\{x_1, \ldots, x_{N_d}\}$, "*Discrete Entropy*" is defined as:

$$H(X_d) = -\sum_{n=1}^{N_d} P(x_n) \log P(x_n) \tag{1}$$

where $P(x_n)$ is the probability that $X_d = x_n$.

Similarly, for a random variable $X_c$ whose value varies continuously, "*Continuous Entropy*" (also called "*Differential Entropy*") is defined as:

$$h(X_c) = -\int_{\Omega} f(x) \log f(x) dx \tag{2}$$

where $f(x)$ is the PDF of $X_c$, and $\Omega$ is the domain over which $X_c$ is defined.

Further, if a continuous variable $X_c$ is converted to (represented as) a discrete variable $X_d$ by binning the continuous values into $N_d$ bins centered at the discrete values $\{x_1, \ldots, x_{N_d}\}$ having bin-width $\Delta$, the relationship between the discrete and continuous values of entropy can (for sufficiently small $\Delta$) be shown to be:

$$\lim_{\Delta \to 0} [H_\Delta(X_d) + \log(\Delta)] = h(X_c) \tag{3}$$

where $H_\Delta(X_d)$ is the discrete entropy with bin width $\Delta$ and $h(X_c)$ is the corresponding continuous entropy. ($H_\Delta(X_d) = -\sum_{n=1}^{N} P(x_n) \log P(x_n)$, $t_n - \frac{1}{2}\Delta \le x_n \le t_n + \frac{1}{2}\Delta$, where $t_n$ is the central point of $n$th bin.) This limit exists if the PDF of random variable X is Riemann integrable; according to the definition of continuous entropy, the value of entropy is determined by the PDF and is independent of bin-width $\Delta$.

Note that discretization happens when we observe (measure/record) values for continuous variables with some finite level of precision $\Delta$, rounding up or down to the nearest finite precision value available to the measuring device. Equation (3) above tells us that the loss of entropy associated with this process is about $\log(\Delta)$ provided that $\Delta$ is sufficiently small. Therefore, equation (3) can be used to estimate the continuous entropy of a variable, by adding $\log(\Delta)$ to the estimate of discrete entropy computed from the (discretely observed) observations. For further details, please see section 8.3 of *Cover and Thomas* [2006]. To avoid misunderstanding, please note that the "precision $\Delta$" and "optimal bin width h" are two different concepts. The precision $\Delta$ is determined by the measurement process, while the optimal bin width h is determined by the derivative of the PDF (as shown in section 3.2.1). So in this paper, the two concepts are discussed separately.

### 3.2. PDF Estimation

PDF estimation methods are of two types: parametric and nonparametric. Parametric methods prespecify (assume) a parametric model/hypothesis for the underlying PDF of the data (such as *Gaussian*, *gamma*, *exponential*, etc.), and then use the data to estimate values for the parameters of the model. Then, known analytical expressions obtained by plugging the PDF into equation (2) and integrating can be used to compute entropy.

Parametric methods are only suitable when the data are known (or can be confidently assumed) to follow the presumed distribution, and can be inadequate or inappropriate in cases where the underlying distribution is unknown or not well approximated by the model hypothesis. For example, *Li et al*. [2012, 2013] have

shown that it may often be necessary to develop hybrid distributions for commonly studied hydrological variables.

To provide a more general approach, we focus here on nonparametric methods that impose fewer assumptions on the shape of the underlying distribution. Below we briefly review three popular nonparametric PDF estimation methods (*Bin-Counting*, *Kernel Density Estimation*, and *Average Shifted Histograms*) that have been commonly used in hydrology for entropy estimation, and also for other purposes such as visualization and statistical analysis.

#### 3.2.1. Bin Counting

The bin-counting (or histogram) method is the oldest and most widely used PDF estimation technique. Suppose that $x_0$ is a give value and $h$ is the bin width (here we use the notation $h$ for bin width to distinguish it from $\Delta$ which we have used earlier to denote measurement precision), the $i$th bin of the histogram is defined as the left-close right-open interval $[x_0+ih, x_0+(i+1)h)$ [*Silverman*, 1986], and the estimated PDF $\hat{f}(x)$ is given by:

$$\hat{f}(x)=\frac{v_i}{N_{Data}h} \qquad for \quad x\in[x_0+ih, x_0+(i+1)h),\ \mathrm{i}=1,\ \ldots\ N_{Bin} \tag{4}$$

where $v_i$ is the number of sample points falling into the $i$th bin, $N_{Bin}$ is the total number of bins, and $N_{Data}$ is the total number of data values (sample points).

The bin-counting method is easy to understand and use, and can be quite robust. However, the bin width parameter $h$ has a significant influence on robustness and accuracy of the estimated PDF. If the bin width is too small (compared to the roughness of the underlying PDF and the sampling density of the data points), the histogram will be too "*rough*," while if bin width is too large the histogram will be oversmoothed. From the perspective of statistics, an overly rough approximation of the PDF means that the estimated PDF of different sample sets generated from the same distribution will be very different, while an overly smooth PDF can result in large bias. *Scott* [1979] showed that the optimal bin width depends on the actual roughness of the underlying density, and derived an "*optimal*" bin width by minimizing an *Integrated Mean Squared Error* criterion (IMSE) designed to balance variance and bias errors. The expression for optimal bin width is:

$$h^*=\left[\frac{6}{N_{Data}R(f')}\right]^{1/3} \tag{5}$$

where $f$ is the true density function, $f'$ is its derivative, and $R(f')=\int f'(x)^2dx$ is the roughness of $f$. In the case of an approximately Gaussian density, the optimal bin width is given by:

$$h^*=2\times3^{1/3}\pi^{1/6}\sigma {N_{Data}}^{-1/3}\approx 3.49\sigma {N_{Data}}^{-1/3} \tag{6}$$

which is called the "*normal reference rule.*" *Scott* [2004] also provided an "*oversmoothed bandwidth*" rule, based on the fact that the smoothest PDF having variance $\sigma^2$ is given by:

$$g(x)=\frac{15}{16\sqrt{7}\sigma}\left(1-\frac{x^2}{7\sigma^2}\right)^2 \quad -\sqrt{7}\sigma<x<\sqrt{7}\sigma \tag{7}$$

and because $R(f')\geq R(g')$ for any other density function, an upper bound on the optimal bin width can be derived as being (with convergence rate of IMSE to zero being $O({N_{Data}}^{-2/3})$):

$$h^*=\left[\frac{6}{N_{Data}R(f')}\right]^{1/3}\leq\left[\frac{6}{N_{Data}R\left(g'\right)}\right]^{\frac{1}{3}}=\left[\frac{686\sigma^3}{5\sqrt{7}N_{Data}}\right]^{\frac{1}{3}}\approx 3.73\sigma N_{Data} \tag{8}$$

For practical applications, if the PDF appears to be relatively smooth, one may use either the "*normal reference rule*" or the "*oversmoothed bandwidth rule*." However, if the PDF appears to be heavily skewed or has very heavy tails a correction factor must be applied [*Scott*, 1979].

### 3.2.2. Kernel Density Estimator

A kernel density estimator with kernel function $K$ is defined as:

$$\hat{f}(x)=\frac{1}{N_{Data}h}\sum_{i=1}^{N_{Data}}K\left(\frac{x-x_i}{h}\right) \tag{9}$$

where $x_i$ are the sample values, $N_{Data}$ is the total number of sample points, $K$ is the selected kernel density function that is required to satisfy $\int_{-\infty}^{\infty}K(x)dx=1$, and $h$ is the tunable smoothing parameter (also called the "*window width*" or "*bandwidth*"). The topic of kernel density estimation was intensively studied from the 1960s to the 1980s, and the selection of kernel functions, optimal smoothing parameters, and many other technical details have been widely discussed [see *Scott*, 2008]. Perhaps, the most frequent choices for kernel function have been the Beta density with finite support on (−1,1) and the Gaussian density with infinite support.

*Scott* [2008] showed that the optimal choice for smoothing parameter of a nonnegative kernel is:

$$h^*=\left[\frac{R(K)}{\sigma_K^4R(f'')}\right]^{1/5}N_{Data}{}^{-1/5} \tag{10}$$

where $\sigma_K^2$ is the variance of the selected kernel density; the convergence rate is $O(N_{Data}{}^{-4/5})$, which is better than that of bin-counting method. Further, if the underlying data PDF is approximately Gaussian and the selected kernel function is also Gaussian, this simplifies to:

$$h=(4/3)^{1/5}\sigma N_{Data}{}^{-1/5}\approx 1.06\sigma N_{Data}{}^{-1/5} \tag{11}$$

### 3.2.3. Average Shifted Histogram

Whereas the bin-counting method results in a *piecewise constant* approximation to the underlying PDF, the kernel density estimator provide a *continuous* approximation. In general, whereas kernel density methods are typically more accurate they are more costly (in terms of computation) than the bin-counting approach. An efficient and effective alternative method that is both computationally quick (similar to the bin-counting method) but approaches the accuracy of the kernel density method is the Average Shifted Histogram (ASH) approach proposed by *Scott* [1985].

In ASH, each histogram bin of width h is partitioned into $N_{SBin}$ subbins having subbin width $\delta=h/N_{SBin}$. Let the $k$th subbin be denoted as $B_k\equiv[k\delta,(k+1)\delta)$ and let $v_k$ be the number of samples in bin $B_k$. Then, the ASH estimate of the PDF is given by:

$$\hat{f}(x)=\frac{1}{N_{Data}h}\sum_{i=1-N_{SBin}}^{N_{SBin}-1}w(i)v_{k+i}\qquad \text{for}\quad x\in B_k \tag{12}$$

where $w(i)$ is a weighting function that satisfies the conditions of symmetry ($w(i)=w(-i)$) and nonnegativity ($w(i)\geq 0$), and the weights sum to one $\left(\sum_{i=1-N_{SBin}}^{N_{SBin}-1}w(i)/N_{SBin}=1\right)$. Further, the general weighing function is defined as:

$$w(i)=\frac{N_{SBin}\times K(i/N_{SBin})}{\sum_{j=1-N_{SBin}}^{N_{SBin}-1}K(j/N_{SBin})}\qquad \text{for}\quad i=1-N_{SBin},\ldots,N_{SBin}-1 \tag{13}$$

where $K(t)$ is a kernel density function defined on (−1,1). A simple choice for this kernel density function is the isosceles triangular kernel $K(t)=1-|t|$ for $|t|<1$, and the corresponding weighting function has the simple form $w(i)=1-|i|/N_{SBin}$. An alternative, smoother, choice for the density function is the biweight (or quartic) kernel, which is quite popular:

$$K(t)=\frac{15}{16}\left(1-t^2\right)^2 \quad for \quad t\in(-1,1) \tag{14}$$

The ASH estimator has three "parameters" to be tuned, being the form of the kernel function, the subbin number $N_{S\,Bin}$, and the coarse bin width $h$. As mentioned earlier, the biweight kernel has been a popular choice for kernel density. Note that if the subbin number $N_{S\,Bin}=1$, the method reduces to the bin-counting method, while as $N_{S\,Bin}\rightarrow\infty$, the method approaches the kernel density estimator. For practical application, a choice of $N_{S\,Bin}=10$ to 15 has been found to be sufficient. This leaves the selection of the coarse bin width $h$ which can be determined by minimizing IMSE. While the general solution is too complex to be obtained analytically, *Scott* [1985] showed that the solution for the simple isosceles triangular kernel is (as $N_{S\,Bin}\rightarrow\infty$) given by:

$$h^*=\left[\frac{24}{N_{Data}R(f'')}\right]^{1/5} \quad \left\{\approx 2.576\sigma {N_{Data}}^{-\frac{1}{5}} \quad normal\ \ reference\ \ rule\right\} \tag{15}$$

And the convergence rate is $O({N_{Data}}^{-1})$. Numerical experiments have shown that the convergence rate of the ASH method is similar to that of the kernel density estimator, and superior to that of the bin-counting method [*Scott*, 2008]. In the context of analyzing hydrologic data, *Fernando et al*. [2009] proposed the following empirical equation for selection of $h$ and $N_{S\,Bin}$:

$$N_{S\,Bin}=67.412\sigma {N_{Data}}^{-0.2376} \qquad h=2.3766\sigma {N_{Data}}^{-0.1505} \tag{16}$$

### 3.3. Strategy for Entropy Estimation for Hydrologic Data

#### 3.3.1. Dealing With the Zero Effect

When there is a large number of zero values in a data set, PDF estimators designed for continuous distributional forms are not appropriate due to the discontinuity in form of the PDF at the origin. In this case, one must deal with the zero and nonzero values separately.

Consider $N_{Data}$ samples of nonnegative random variable $X$ drawn from a continuous PDF $f(x)$ where $N_x$ is the number of nonzero values, so that the nonzero ratio $k_x=N_x/N_{Data}$. If $[0,+\infty)$ is discretized into $N_{Bin}$ bins, each with bin width $h_i$, the coordinates of bin edges can be expressed as:

$$\{t_i, i=1,2,\ldots,N_{Bin}+1\} \tag{17}$$

In which $t_1=0, \quad t_{i+1}=t_i+h_i, \quad i=1,2,\ldots,N_{Bin}$. The PDF can be expressed as:

$$\begin{cases} P(X=t_1)=1-k_x, & i=1 \\ P(t_i<X\le t_{i+1})=k_x f(x_i)h_i, & i>1 \end{cases} \tag{18}$$

where $x_i$ satisfies $t_i<x_i\le t_{i+1}$ and $f(x_i)h_i=\int_{t_i}^{t_{i+1}}f(x)dx$. According to the First Mean Value Theorem for Integrals, $x_i$ always exist.

Substituting (18) into equation (1), we can compute the entropy $H(X)$ as:

$$H(X)=-(1-k_x)\log(1-k_x)-k_x\log(k_x)-k_x\sum_{i=1}^{N_{Bin}}f(x_i)\ \log(f(x_i))\ h_i-\sum_{i=1}^{N_{Bin}}k_x f(x_i)\ h_i\ \log(h_i) \tag{19}$$

Now consider the four component terms:

$$\mathrm{H1}=-(1-k_x)\log(1-k_x) \tag{20}$$

$$\mathrm{H2}=-k_x\log k_x \tag{21}$$

$$\mathrm{H3} = -k_x \sum_{i=1}^{N_{Bin}} f(x_i)\ \log\left(f(x_i)\right)\ h_i \tag{22}$$

$$\mathrm{H4} = -\sum_{i=1}^{N_{Bin}} k_x f(x_i)\ h_i\ \log\left(h_i\right) \tag{23}$$

Note that terms H1 and H2 are easily computed constant values that depend only on the nonzero ratio $k_x$. The term H3 converges to the differential entropy of $f(x)$ for $x > 0$, as follows:

$$\mathrm{H3} = -k_x \sum_{i=1}^{N_{Bin}} f(x_i)\ \log\left(f(x_i)\right)\ h_i \approx -k_x \int_0^{\infty} f(x)\log f(x)dx \quad \text{as } h_i \to 0 \tag{24}$$

Finally, considering the assumption that the bin width $h_i$ does not change with $x$, term H4 can be simplified to:

$$\mathrm{H4} = -k_x \log\left(h\right) \sum_{i=1}^{N_{Bin}} f(x_i)\ h = -k_x \log\left(h\right) \tag{25}$$

With this background, we can compute the discrete entropy of a data set having many zero values as the sum of the four terms. Terms H1 and H2 are constants depending directly on the nonzero ratio $k_x$. Term H3 is independent of bin width $h$ for sufficiently small $h$, and can be estimated by fitting a PDF to the nonzero data values using one of the methods discussed earlier. Term H4 can be understood as the "*discrete term*." Remember that since discrete entropy can be transformed to differential entropy by removing the "*discrete term*," the differential entropy can also be computed simply by summing terms H1, H2, and H3.

### 3.3.2. Dealing With Highly Skewed Data

Data having severely skewed PDFs are very common in hydrology. In such cases, the use of bin widths or smoothing parameters that are uniform (constant) across the range of the data can introduce significant errors into the estimation process [*Scott*, 1979]. One relatively simple way to address this problem is to transform the original data to form that is less skewed so that the uniform optimal bin width approach can be applied. Once the differential entropy of $h(X)$ is computed in the transformed space, the entropy $h(Y)$ in the original space can be computed by noting that the relationship between the PDFs $f_x(x)$ of $X$ and $f_y(y)$ of $Y$ is, for monotonic transformations, given by $f_y(y) = |g'(y)| f_x(x)$ [*Ross*, 2010, section 2.5.4]. Further, if $Y > 0$, then $g'(Y) > 0$, we have $f_y(y) = g'(y) f_x(x)$.

Applying the definition for differential entropy, we can derive the general expression relating entropy before and after transformation as:

$$h(Y) = h(X) - \int_0^{\infty} f_x(x)\log g'(y)dx \tag{26}$$

In the hydrological literature, it is common to transform non-Gaussian data having heteroscedastic measurement precision to a form that is approximately Gaussian and homoscedastic using the Box-Cox power function [*Box and Cox*, 1982]:

$$x = B(y) = \begin{cases} \dfrac{(y+1)^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ \log\left(y+1\right) & \lambda = 0 \end{cases} \tag{27}$$

where nonnegative values of the original value $y$ are transformed to nonnegative value $x$, and $\lambda$ is a parameter controlling the strength of the transformation. Note that *Chapman* [1986] investigated only the special

case of the log transformation which corresponds to $\lambda=0$. Here we present the more general case where $\lambda$ can take on other values. Since the expression for the derivative of the Box-Cox transformation is:

$$g'(y)=\begin{cases}(y+1)^{\lambda-1} & \lambda\neq 0\\ \dfrac{1}{y+1} & \lambda=0\end{cases} \tag{28}$$

we obtain the result:

$$h(Y)=\begin{cases}h(X)+\dfrac{1-\lambda}{\lambda}\displaystyle\int f_x(x)\log(\lambda x+1)dx & \lambda\neq 0\\ h(X)+E(X) & \lambda=0\end{cases} \tag{29}$$

Equation (29) indicates that, under transformation of the data, the term H3 derived above consists of two parts H3a and H3b, the first being the entropy of the transformed variable $X$ and the other being a correction term associated with the transformation (note that the transformation is only applied to the nonzero portion of the data):

$$H3a=k_x h(X)=-k_x\int_0^\infty f_x(x)\log f_x(x)dx \tag{30}$$

$$H3b=\begin{cases}k_x\dfrac{1-\lambda}{\lambda}\displaystyle\int f_x(x)\log(\lambda x+1)dx & \lambda\neq 0\\ k_x E(X) & \lambda=0\end{cases} \tag{31}$$

From equation (31), we see that term H3b is equal to zero when $\lambda=1$, which corresponds to no transformation.

#### 3.3.3. Dealing With the Measurement Effect

As discussed in section 2, heteroscedasticity in observation error can be reduced by use of the Box-Cox transformation, which simultaneously reduces the influence of high skewness. To deal with the error caused by the minimum measurement unit, we must be careful about two issues. (1): to compute differential entropy, we need only compute the sum of terms H1, H2, and H3, while if discrete entropy is required, we can also add term H4 where $h$ is chosen to be equal to the precision $\Delta$ of the measurement system. (2) When computing term H3, the smoothing parameters must be chosen with care. Note that in the binning methods, the bin width $h$ (and also the subbin width $\delta$) must be *larger* than the precision $\Delta$ of the measurement system. If the kernel density estimator is used, the smoothing parameter must also be carefully selected; to be prudent, the *variance* $\sigma_K^2$ of the selected kernel density function should be larger than the precision $\Delta$ of the measurement system. In our experience, the optimal bin width is usually larger than the minimum measurement unit (precision $\Delta$). However, the measurement effect may be amplified by Box-Cox transformation. For example, if the minimum measurement unit is 0.1 mm, and the transformation parameter $\lambda=0$, the distance between (0 mm, 0.1 mm] is extended to 0.0953, while the distance between (99.9 mm, 100 mm] is only $9.9059\times10^{-4}$. So rigorous checking of the bin width is necessary for low values of rainfall and runoff if the Box-Cox transformation is used.

## 4. Case Studies

### 4.1. Case 1: Synthetically Generated Data

To assess the accuracy of the method proposed above, we estimate the entropy of data generated synthetically from two discrete-continuous hybrid distributions for which the entropy can be analytically computed (see Table 1). The nonzero portions of the data are generated using *lognormal* and *gamma* distributions, both of which are positively skewed.

**Table 1.** Function Forms of the Lognormal and Gamma Distributions, and the Theoretical Value of *Entropy* for Each (Cited From Table 17.1 of *Cover and Thomas* [2006])[a]

| Name | Density | Entropy (nats) |
|---|---|---|
| Lognormal | $f(x)=\frac{1}{\sigma x\sqrt{2\pi}}e^{-\frac{[\ln(x)-\mu]^2}{2\sigma^2}}$ | $\mu+\frac{1}{2}\ln(2\pi e\sigma^2)$ |
| Gamma | $f(x)=\frac{1}{b^a\Gamma(a)}x^{a-1}e^{-x/b}$ | $\ln(b\Gamma(a))+(1-a)\psi(a)+a$ |

[a]Entropy is computed using logarithms to the base e and reported in "nats."

For the lognormal distribution, the variable $x > 0$ and the parameters $\mu$ and $\sigma > 0$ all are real numbers. If variable $u \sim N(\mu, \sigma^2)$ where $\mu$ and $\sigma$ are the mean and standard deviation of the normal distribution, and $u=\ln x$, then $x$ is distributed as lognormal denoted by $x \sim \Lambda(\mu, \sigma^2)$. For the gamma distribution, parameters $a$ and $b$ are positive real numbers, where $a$ is the shape parameter, $b$ is the scale parameter, $\Gamma(.)$ is the gamma function, and $\psi(.)$ is the digamma function.

Note that terms H1 (equation (20)) and H2 (equation (21)) depend only on $k_x$, the number of zero values in the data set. Term H3 (equation (22)) represents the differential entropy of the nonzero data values, which can be analytically derived for the two distributions considered here:

$$H3=\begin{cases}-k_x\left[\mu+\frac{1}{2}\ln\left(2\pi e\sigma^2\right)\right] & \textit{for lognormal distribution}\\ -k_x[\ln(b\Gamma(a))+(1-a)\psi(a)+a] & \textit{for Gamma distribution}\end{cases} \quad (32)$$

For each synthetic data set case, we consider a sample length of $N_{Data}=10,000$ values (representing nearly 30 years of daily data), of which one half are set to be zero values (so the nonzero fraction $k_x=0.5$). For the *Lognormal* case, we set the parameters to be $\mu=1$ and $\sigma=1$ and for the gamma distribution we set the parameters to be $a=3$ and $b=0.5$. These correspond to skewed distributions as illustrated in Figure 2.

To mimic the "*discrete effect*" caused by available precision in the measurement process, we set the precision to $\Delta=0.1$ and round the value of the synthetically generated values up or down to the closest discrete value. Therefore, the discrete term $H4=-k_x log\Delta \approx 1.1513$. As discussed above, term *H4* depends only on the measurement precision and does not depend on the choice of method for estimating the PDF. Accordingly, term *H4* is not relevant to the following intercomparisons.

We then compare values of entropy estimated by different approaches with the theoretical values. (1) In the first approach, we directly apply *Bin Counting* (equation (4)) without special consideration of the zero effect or applying any kind of transformation to the data (called "Crude-Bin" in the following tables and figures). (2) In the second, we again apply *Bin Counting* while accounting for the *zero effect* (section 3.3.1, equations (20–22)) but without transformation (labeled as "*Delta-Bin*"). (3) In the third, we apply *Bin Counting* while accounting for the *zero effect* and applying a Box-Cox transformation (equations (27), (30), and (31)) to the data using $\lambda = 0$ (labeled as "*Delta-BoxCox-Bin*"). (4) The fourth approach is similar to the third but using the *Kernel* method (equation (9)) for PDF estimation (labeled as "*Kernel*"). (5) Similarly, in the fifth and sixth, the *ASH* method (equation (12)) is used for PDF estimation (labeled as "*ASH*"). In the fifth, we use finer bin number $N_{SBin} = 10$ while in the sixth we use $N_{SBin} = 3$.

The results are presented in Table 2 and Figure 2. To assess the uncertainty in the estimates, we use the bootstrap method [*Efron*, 1979] implemented in the MATLAB toolbox provided by *Zoubir and Boashash* [1998]; 10,000 bootstrap replications were generated. Because the distribution of the estimates is approximately normal, we report only the mean and 3 times the standard deviation (corresponding to 99.7% confidence intervals). Figure 2 shows the estimated PDFs for one of the replicates for each of the different methods. The results clearly indicate that the *Delta-BoxCox-Bin* approach, *ASH* with $N_{SBin} = 3$ and *kernel* approach provide the best performance, followed by *Delta-Bin*. Note, especially, that the performance of the *Crude-Bin* method and *ASH* with $N_{SBin} = 10$ can be quite poor, especially for the gamma distribution.

From the individual terms listed in Table 2, we can see where the main problems lie. In the case of the *Crude-Bin* method, we do not actually have explicit values for each term, so we have used the value of entropy computed from the "*first*" bin (the one containing x = 0) as term H1 and the sum of remaining bins as H2 + H3. While the values of H2 + H3 are close to the theoretical values, the value of H1 for the *Crude-Bin* method is much smaller than it should be, indicating that errors caused by the "zero effect" can be quite large if not properly handled. Further, Figure 2b shows that if we try to estimate the PDF without first

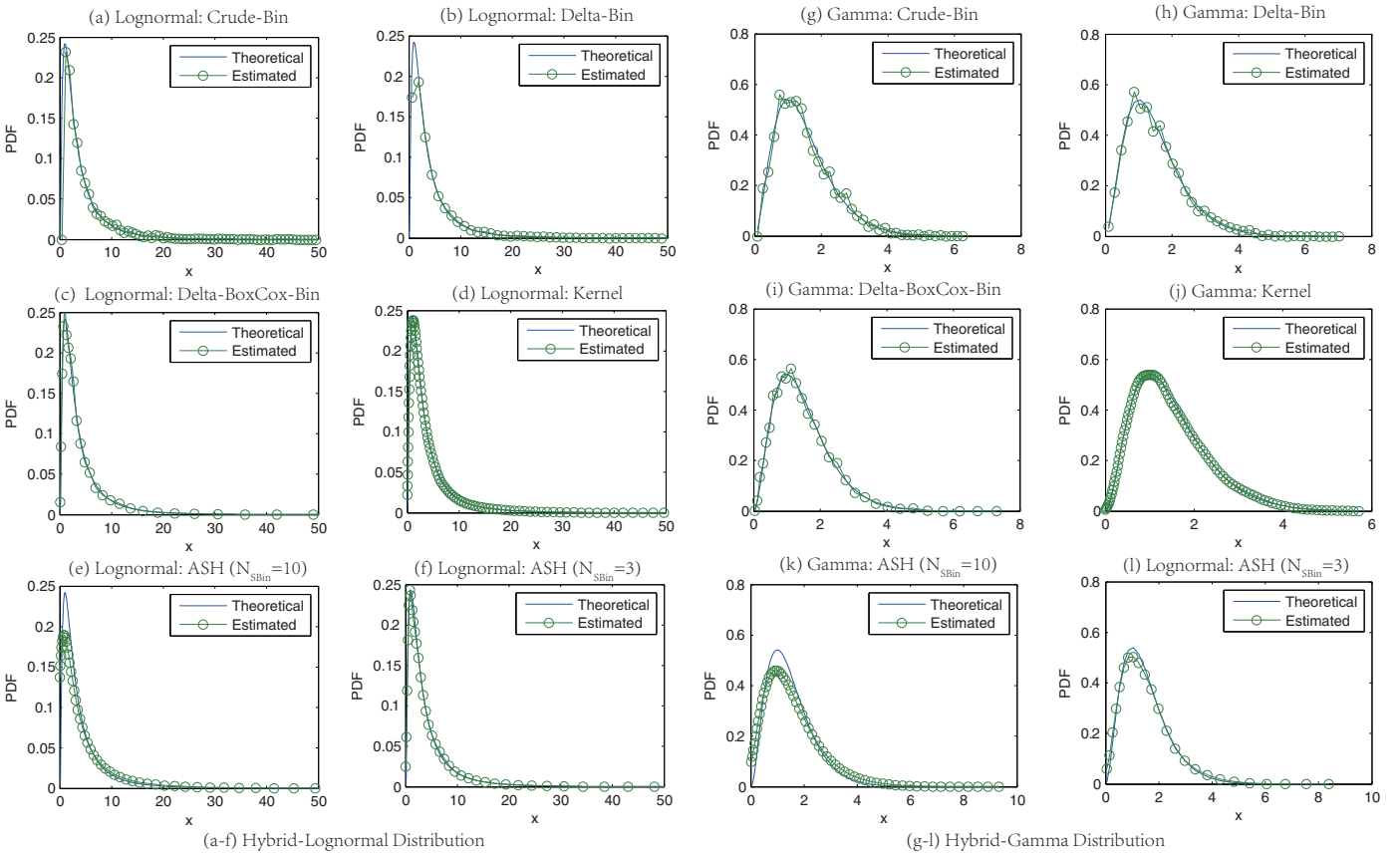


**Figure 2.** Estimated and theoretical PDFs of the nonzero part for the (left) hybrid-lognormal and (right) hybrid-gamma distributions. (a) Lognormal distribution, Crude-Bin method, no delta function, no Box-Cox transformation; (b) lognormal distribution, Delta-Bin method, use delta function, no Box-Cox transformation; (c) lognormal distribution, Delta-BoxCox-Bin method, use delta function, and Box-Cox transformation; (d) lognormal distribution, Kernel method, use delta function, and Box-Cox transformation; (e) lognormal distribution, ASH method, use delta function, and Box-Cox transformation, finer bin width $N_{SBin} = 10$; (f) lognormal distribution, ASH method, use delta function, and Box-Cox transformation, finer bin width $N_{SBin} = 3$; (g) gamma distribution, Crude-Bin method; (h) gamma distribution, Delta-Bin method; (i) gamma distribution, Delta-BoxCox-Bin method; (j) gamma distribution, Kernel method; (k) gamma distribution, ASH method, $N_{SBin} = 10$; (l) gamma distribution, ASH method, $N_{SBin} = 3$.

applying a transformation we can incur significant error in the region of the peak due to the peak being oversmoothed. These results confirm *Scott* [1979] that uniform bin widths should not be used in PDF estimation if the distribution is seriously skewed, and the transformation can help to mitigate this problem.

**Table 2.** Theoretical and Estimated Entropy for the Hybrid-Lognormal and Hybrid-Gamma Distributions Computed From 10,000 Synthetically Generated Samples[a]

| | Theoretical | Crude-Bin | Delta-Bin | Delta-BoxCox-Bin | Kernel | ASH ($N_{SBin} = 10$) | ASH ($N_{SBin} = 3$) |
|---|---|---|---|---|---|---|---|
| *Hybrid Lognormal* | | | | | | | |
| Entropy | 1.9026 | 1.6554 ± 0.0754[b] | 1.9350 ± 0.0453 | 1.8938 ± 0.0436 | 1.9329 ± 0.0441 | 1.9787 ± 0.0459[b] | 1.8858 ± 0.0532 |
| H1 | 0.3466 | 0.2254 ± 0.0704[b] | 0.3464 ± 0.0046 | 0.3471 ± 0.0046 | 0.3484 ± 0.0044 | 0.3458 ± 0.0047 | 0.3450 ± 0.0057 |
| H2+H3 | 1.5661 | 1.4295 ± 0.0424[b] | 1.5887 ± 0.0416 | 1.5467 ± 0.0398 | 1.5845 ± 0.0406 | 1.6329 ± 0.0420[b] | 1.5410 ± 0.0488 |
| H2 | 0.3466 | | 0.3468 ± 0.0045 | 0.3461 ± 0.0047 | 0.3447 ± 0.0048 | 0.3473 ± 0.0045 | 0.3481 ± 0.0053 |
| H3 | 1.2095 | | 1.2419 ± 0.0450 | 1.2007 ± 0.0436 | 1.2398 ± 0.0444 | 1.2855 ± 0.0458[b] | 1.1927 ± 0.0529 |
| $N_{Bin}$ | | 96 | 61 | 27 | | 50 | 39 |
| *Hybrid Gamma* | | | | | | | |
| Entropy | 1.2704 | 0.3396 ± 0.0573[b] | 1.2654 ± 0.0244 | 1.2689 ± 0.0238 | 1.2771 ± 0.0235 | 1.5221 ± 0.0264[b] | 1.2888 ± 0.0235 |
| H1 | 0.3466 | −0.5560 ± 0.0379[b] | 0.3480 ± 0.0044 | 0.3480 ± 0.0044 | 0.3466 ± 0.0046 | 0.3472 ± 0.0045 | 0.3443 ± 0.0048 |
| H2+H3 | 0.9238 | 0.8956 ± 0.0224[b] | 0.9173 ± 0.0218 | 0.9209 ± 0.0209 | 0.9304 ± 0.0204 | 1.1748 ± 0.0223[b] | 0.9445 ± 0.0199 |
| H2 | 0.3466 | | 0.3451 ± 0.0047 | 0.3451 ± 0.0047 | 0.3465 ± 0.0046 | 0.3459 ± 0.0047 | 0.3487 ± 0.0044 |
| H3 | 0.5772 | | 0.5723 ± 0.0246 | 0.5759 ± 0.0240 | 0.5840 ± 0.0235 | 0.8289 ± 0.0265[b] | 0.5958 ± 0.0231 |
| $N_{Bin}$ | | 46 | 34 | 22 | | 30 | 21 |

[a]Unit: nats. Uncertainties correspond to 3 standard deviations (99.7% confidence intervals). $N_{Bin}$ represents the medium number of bins.
[b]Estimates that are significantly biased (confidence intervals excludes the true values).

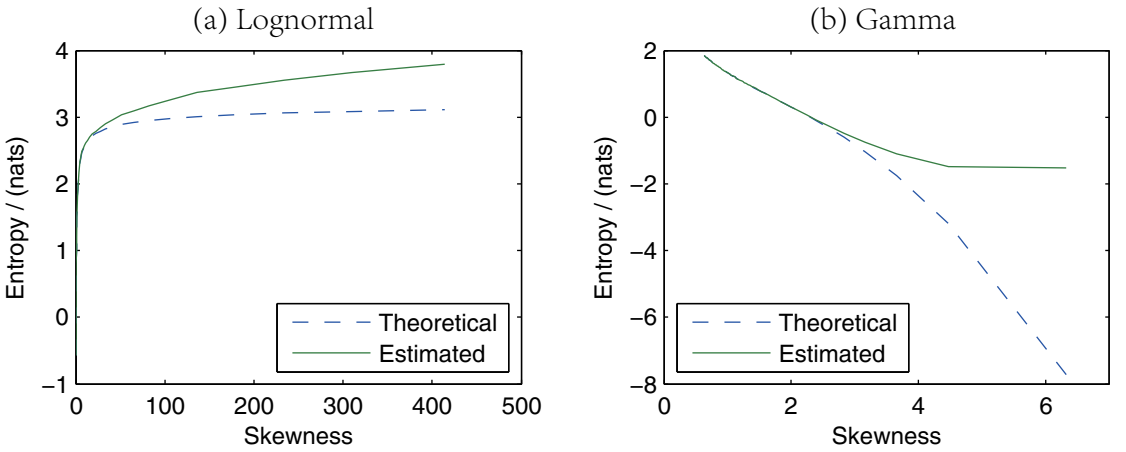


**Figure 3.** The relationship between skewness and entropy estimation error. (a) Lognormal distribution with $\mu$=1 and varying $\sigma$ from 0.05 to 2; (b) gamma distribution with $b$=0.5 and varying $a$ from 0.1 to 10.

In Table 2 and Figure 2, we also compared different finer bin number $N_{SBin}$. Theoretically for large $N_{SBin}$ the ASH estimator converge to Kernel estimator [*Scott*, 1985], but in fact large $N_{SBin}$ maybe not a good choice. As shown in Figures 2e and 2k, with $N_{SBin}$ = 10 the peak of PDF is oversmoothed due to the fact that ASH's

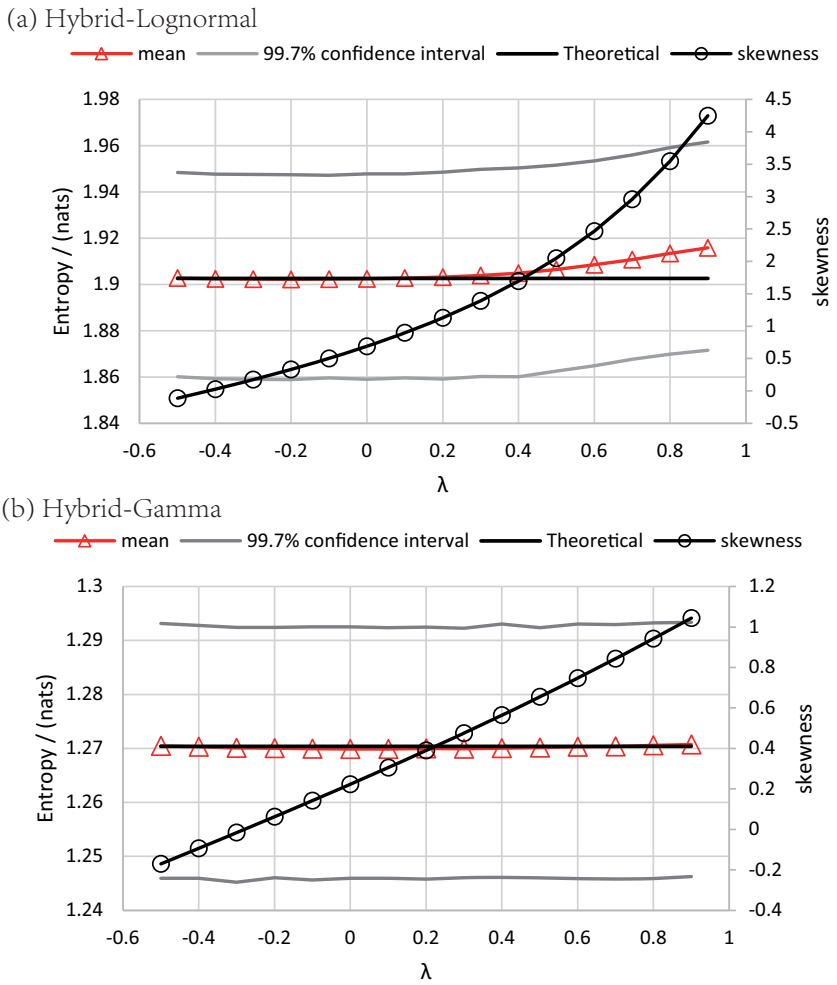


**Figure 4.** The influence of Box-Cox transformation parameter $\lambda$ to the estimated entropy. (a) Hybrid-lognormal; (b) hybrid-gamma.

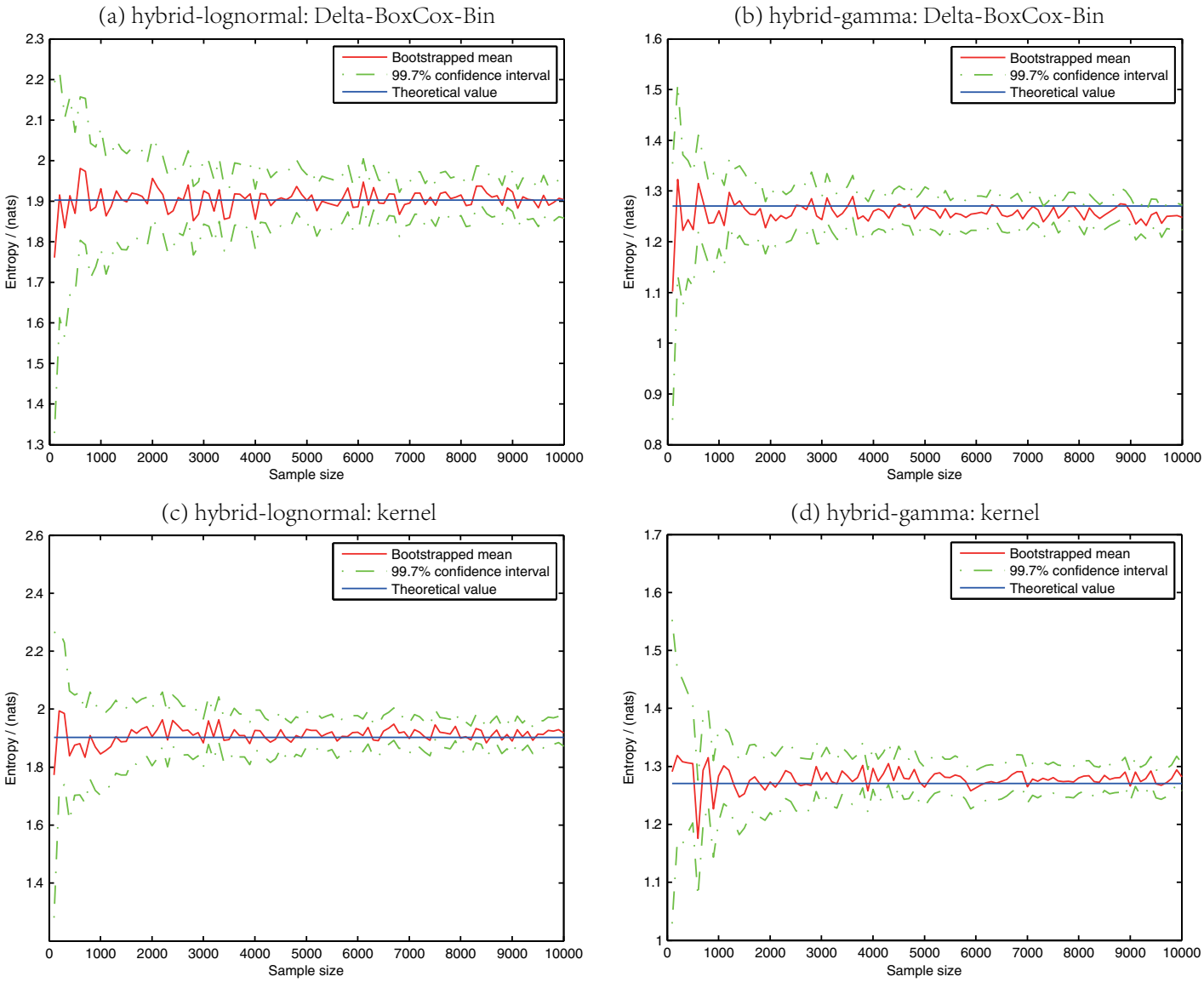


**Figure 5.** The influence of sample size on the accuracy, precision, and robustness of the entropy estimates; uncertainties correspond to 3 standard deviations (99.7% confidence intervals). (a) Hybrid-gamma distribution, using "*Delta-BoxCox-Bin*" method; (b) hybrid-lognormal distribution, using "*Delta-BoxCox-Bin*" method; (c) hybrid-gamma distribution, using "*kernel*" method; (d) hybrid-lognormal distribution, using "*kernel*" method.

mechanism involves a weighted average of nearby finer bins. Consequently $N_{SBin}$ needs to be fine tuned to guarantee good performance, such as the cases of $N_{SBin} = 3$ in Figures 2f and 2l. For practical observation data, we do not known the "true" distribution as a reference for tuning, so *Bin-Counting* and *Kernel* methods might be more generally appropriate choice, being both relatively accurate and robust.

There is another issue of choosing appropriate parameter $\lambda$ for Box-Cox transformation. First, we investigate the influence of skewness to entropy estimation error. We use two synthetic cases: (a) lognormal distribution with $\mu=1$ and varying $\sigma$ from 0.05 to 2; (b) gamma distribution with $b=0.5$ and varying $a$ from 0.1 to 10. As shown in Figure 3, the estimated entropy is positively biased if skewness is larger than 2, indicating the necessity of reducing skewness with Box-Cox transformation. In Figure 4, we use the hybrid distributions and tried $\lambda$ from $-0.5$ to 0.9. For both of the cases, the entropy estimation error is reduced to less than 0.01 nats if skewness is transformed to less than 2.

To assess the influence of sample size on the accuracy and robustness of the entropy estimates, Figure 5 presents results for sample sizes: 100 (one season) to 10,000 (~30 years). The methods used are "*Delta-*

**Table 3.** Entropy of Daily Precipitation and Runoff of the Leaf River Computed From a 40 Year Data Set[a]

| Years | Annual Precipitation (mm) | Ratio of Rainy Days—$k_x$ | Entropy of Rainy Days (nats) | Total Entropy of Precipitation (nats) | Annual Runoff Depth (mm) | Entropy of Daily Runoff Depth (nats) |
|---|---|---|---|---|---|---|
| 1948–1988 (40 year) | 1432 | 0.4511 ± 0.0124 | 1.3054 ± 0.0462 | 1.9938 ± 0.0481 | 501.61 | 0.9486 ± 0.0435 |
| 1948–1958 (10 year) | 1370 | 0.4359 ± 0.0246 | 1.2933 ± 0.0892 | 1.9781 ± 0.0944 | 451.45 | 0.7936 ± 0.0924 |
| 1958–1968 (10 year) | 1350 | 0.4435 ± 0.0247 | 1.2843 ± 0.0899 | 1.9710 ± 0.0945 | 454.56 | 0.7678 ± 0.0877 |
| 1968–1978 (10 year) | 1521 | 0.4674 ± 0.0247 | 1.3667 ± 0.0916 | 2.0575 ± 0.0941 | 565.01 | 1.1251 ± 0.0810 |
| 1978–1988 (10 year) | 1486 | 0.4572 ± 0.0249 | 1.3197 ± 0.0922 | 2.0092 ± 0.0955 | 535.43 | 0.9266 ± 0.0884 |

[a]Uncertainties correspond to 3 standard deviations (99.7% confidence intervals).

*BoxCox-Bin"* and "*kernel*," and the number of bootstrap replicates is 10,000. As shown, larger sample sizes can significantly reduce the bias and variance of the estimates; however in all cases the 99.7% confidence intervals encompass the true values. Note also that the shape of distribution has an influence on the width of confidence interval; i.e., the confidence interval of estimates associated with the hybrid-lognormal distribution is larger than that of hybrid-gamma distribution.

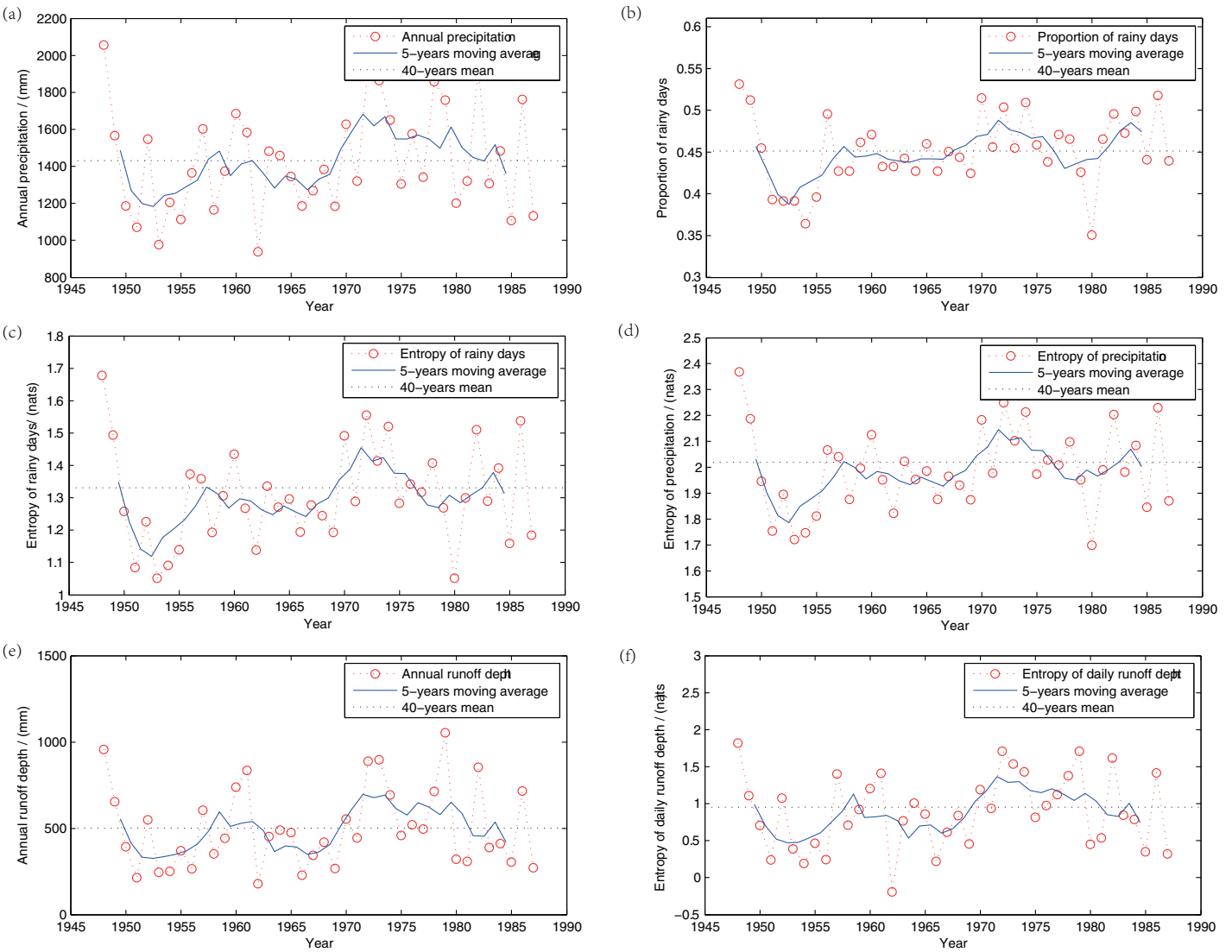


**Figure 6.** Differential entropy of daily precipitation and discharge of Leaf River basin. (a) The annual precipitation. 5 years moving average and 40 years mean value are also shown in the same figure; (b) the ratio of rainy days; (c) the entropy of rainy days (entropy of nonzero values in the data set, namely, H3 in equation (21)); (d) total entropy of precipitation; (e) the annual runoff; (f) entropy of daily runoff depth.

### 4.2. Case 2: Real Hydrological Data

We apply the entropy estimation methods to a widely used 40 year (October 1948 to September 1988) daily time step data set from the 1944 $km^2$ humid Leaf River basin located north of Collins, Mississippi, USA [see, e.g., *Sorooshian et al*., 1993; *Thiemann et al*., 2001; *Vrugt et al*., 2003; *Gong et al*., 2013]. Average annual rainfall for this basin is 1492 mm, and the elevation varies from 60 to 185m. To assess the uncertainty in the estimates, we also use the bootstrap method. Because the 10,000 replicates were found to conform to approximately normal distributions, we report only the mean and 3 times the deviation (corresponding to 99.7% confidence intervals).

The methodology applied was *Delta-BoxCox-Bin*, and to be consistent with previous research [*Bulygina and Gupta*, 2010; *Gong et al*., 2013; *Misirli et al*., 2003], we used $\lambda = 0.3$ for the Box-Cox transformation parameter. Results for daily precipitation and runoff depth are shown in Table 3 for the entire 40 years, and for each of the four successive 10 year periods. Figure 6, shows values computed for each of the 40 years individually, along with their 5 year moving average. This result confirmed that the methods presented in this paper are valid and robust for true observed rainfall and runoff data. Of course, the generality of this result would need to be evaluated using extensive regional or global data sets.

## 5. Discussion and Conclusions

We have discussed a method for computing robust and accurate estimates of entropy that accounts for several important characteristics of hydrological data sets (*zero effect*, *measurement effect*, and *skewness effect*). We have also discussed the relevance of bin-width selection in the context of this problem, and compiled together all of the necessary technical details into one paper so as to facilitate practical applications. The methodology provides a foundation for the estimation of other information-based metrics, such as KL-divergence, mutual information, transfer entropy, etc. Future work is needed to develop methods for obtaining robust estimates of entropy for higher-dimensional joint PDF's.

Our synthetic case study demonstrates clearly that proper handling of zero values is necessary when there is a significant proportion of zero values, and that a Box-Cox transformation can help to mitigate the effects of high levels of skewness. Further, the *Bin-Counting* and *Kernel-Density* methods, applied with optimal bin width selection, are able to provide accurate, precise, and robust estimates of the underlying PDF, and hence of entropy. Our methodology is supported by the real-data Leaf River basin (Mississippi, USA) study.

As always, we invite discussion and collaborations on this and other related topics. The computer code used in this work is available from the first author.

**Acknowledgments**

Support for this study was provided by the National Science Foundation of China (contracts 51025931, 50939004, and 51309011). We would like to acknowledge the thorough and constructive suggestions from Steven Weijs, Nick van de Giesen, and another anonymous referee. Their comments significantly improved the quality of this paper.

## References

Anderson, M. G., and J. J. Mcdonnell (2005), *Encyclopedia of Hydrological Sciences*, John Wiley, Chichester, U. K.

Box, G., and D. R. Cox (1982), An analysis of transformations revisited, rebutted, *J. Am. Stat. Assoc.*, *77*(377), 209–210, doi:10.2307/2287791.

Bulygina, N., and H. Gupta (2010), How Bayesian data assimilation can be used to estimate the mathematical structure of a model, *Stochastic Environ. Res. Risk Assess.*, *24*(6SI), 925–937, doi:10.1007/s00477-010-0387-y.

Chapman, T. G. (1986), Entropy as a measure of hydrologic data uncertainty and model performance, *J. Hydrol.*, *85*(1-2), 111–126, doi:10.1016/0022-1694(86)90079-X.

Cover, T. M., and J. A. Thomas (2006), *Elements of Information Theory*, John Wiley, Hoboken, N. J.

Efron, B. (1979), 1977 Rietz lecture—Bootstrap methods—Another look at the Jackknife, *Ann. Stat.*, *7*(1), 1–26, doi:10.1214/aos/1176344552.

Fernando, T., H. R. Maier, and G. C. Dandy (2009), Selection of input variables for data driven models: An average shifted histogram partial mutual information estimator approach, *J. Hydrol.*, *367*(3-4), 165–176, doi:10.1016/j.jhydrol.2008.10.019.

Gong, W., H. V. Gupta, D. W. Yang, K. Sricharan, and A. O. Hero (2013), Estimating epistemic and aleatory uncertainty during hydrologic modeling: An information theoretic approach, *Water Resour. Res.*, *49*, 2253–2273, doi:10.1002/wrcr.20161.

Hyvarinen, A., and E. Oja (2000), Independent component analysis: Algorithms and applications, *Neural Networks*, *13*(4-5), 411–430, doi:10.1016/S0893-6080(00)00026-5.

Li, C., V. P. Singh, and A. K. Mishra (2012), Simulation of the entire range of daily precipitation using a hybrid probability distribution, *Water Resour. Res.*, *48*, W3521, doi:10.1029/2011WR011446.

Li, C., V. P. Singh, and A. K. Mishra (2013), A bivariate mixed distribution with a heavy-tailed component and its application to single-site daily rainfall simulation, *Water Resour. Res.*, *49*, 767–789, doi:10.1002/wrcr.20063.

May, R. J., H. R. Maier, G. C. Dandy, and T. Fernando (2008), Non-linear variable selection for artificial neural networks using partial mutual information, *Environ. Modell. Software*, *23*(10-11), 1312–1326, doi:10.1016/j.envsoft.2008.03.007.

Misirli, F., H. V. Gupta, S. Sorooshian, and M. Thiemann (2003), Bayesian recursive estimation of parameter and output uncertainty for watershed models, in *Calibration of Watershed Models, Water Sci. Appl. Ser.*, vol 6, edited by Q. Duan, pp. 113–124, AGU, Washington, D. C.

Pechlivanidis, I. G., B. M. Jackson, H. K. Mcmillan, and H. V. Gupta (2012a), Use of informational entropy-based metrics to drive model parameter identification, in *Proceedings of the 12th International Conference on Enviornmental Science and Technology*, edited by Lekkas T. D., Rhodes, Greece, A1476–A1483, Curran Associates, Inc., N. Y.

Pechlivanidis, I. G., B. M. Jackson, H. K. Mcmillan, and H. V. Gupta (2012b), Using an informational entropy-based metric as a diagnostic of flow duration to drive model parameter identification, *Global NEST J.*, *14*(3), 325–334.

Ross, S. M. (2010), *Introduction to Probability Models*, 10th ed., Academic, Boston, Mass.

Ruddell, B. L., and P. Kumar (2009a), Ecohydrologic process networks: 1. Identification, *Water Resour. Res.*, *45*, W03419, doi:10.1029/2008WR007279.

Ruddell, B. L., and P. Kumar (2009b), Ecohydrologic process networks: 2. Analysis and characterization, *Water Resour. Res.*, *45*, W03420, doi:10.1029/2008WR007280.

Schoups, G., and J. A. Vrugt (2010), A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, *Water Resour. Res.*, *46*, W10531, doi:10.1029/2009WR008933.

Scott, D. W. (1979), Optimal and data-based histograms, *Biometrika*, *66*(3), 605–610, doi:10.2307/2335182.

Scott, D. W. (1985), Averaged shifted histograms—Effective nonparametric density estimators in several dimensions, *Ann. Stat.*, *13*(3), 1024–1040, doi:10.1214/aos/1176349654.

Scott, D. W. (2004), *Handbook of Computational Statistics—Concepts and Methods*, Springer, N. Y.

Scott, D. W. (2008), *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley, N. Y.

Scott, D. W., and J. R. Thompson (1983), Probability density estimation in higher dimensions, in *Computer Science and Statistics: Proceedings of the Fifteenth Symposium on the Interface*, edited by J. E. Gentle, 173–179, North-Holland, Amsterdam, Netherlands.

Shannon, C. E. (1948), A mathemetical theory of communication, *Bell Syst. Tech. J.*, *27*(379-423), 623–656.

Sharma, A. (2000a), Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1—A strategy for system predictor identification, *J. Hydrol.*, *239*(1-4), 232–239, doi:10.1016/S0022-1694(00)00346-2.

Sharma, A. (2000b), Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 3—A nonparametric probabilistic forecast model, *J. Hydrol.*, *239*(1-4), 249–258, doi:10.1016/S0022-1694(00)00348-6.

Sharma, A., K. C. Luk, I. Cordery, and U. Lall (2000), Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 2—Predictor identification of quarterly rainfall using ocean-atmosphere information, *J. Hydrol.*, *239*(1-4), 240–248, doi:10.1016/S0022-1694(00)00347-4.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, N. Y.

Singh, V. P. (1997), The use of entropy in hydrology and water resources, *Hydrol. Processes*, *11*(6), 587–626, doi:10.1002/(SICI)1099-1085(199705)11:6<587::AID-HYP479>3.3.CO;2-G.

Singh, V. P. (2000), The entropy theory as a tool for modelling and decision-making in environmental and water resources, *Water SA*, *26*(1), 1–11.

Singh, V. P. (2013), *Entropy Theory and Its Application in Environmental and Water Engineering*, Wiley-Blackwell, Oxford, U. K.

Sorooshian, S., and J. A. Dracup (1980), Stochastic parameter-estimation procedures for hydrologic rainfall-runoff models—Correlated and heteroscedastic error cases, *Water Resour. Res.*, *16*(2), 430–442, doi:10.1029/WR016i002p00430.

Sorooshian, S., Q. Y. Duan, and V. K. Gupta (1993), Calibration of rainfall-runoff models—Application of global optimization to the sacramento soil-moisture accouting model, *Water Resour. Res.*, *29*(4), 1185–1194, doi:10.1029/92WR02617.

Thiemann, M., M. Trosset, H. Gupta, and S. Sorooshian (2001), Bayesian recursive parameter estimation for hydrologic models, *Water Resour. Res.*, *37*(10), 2521–2535, doi:10.1029/2000WR900405.

Vrugt, J. A., H. V. Gupta, W. Bouten, and S. Sorooshian (2003), A shuffled complex evolution metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters, *Water Resour. Res.*, *39*(8), 1201, doi:10.1029/2002WR001642.

Weijs, S. V., and N. van de Giesen (2011), Accounting for observational uncertainty in forecast verification: An information-theoretical view on forecasts, observations, and truth B-5010-2008, *Mon. Weather Rev.*, *139*(7), 2156–2162, doi:10.1175/2011MWR3573.1.

Weijs, S. V., R. van Nooijen, and N. van de Giesen (2010a), Kullback-Leibler divergence as a forecast skill score with classic reliability-resolution-uncertainty decomposition, *Mon. Weather Rev.*, *138*(9), 3387–3399, doi:10.1175/2010MWR3229.1.

Weijs, S. V., G. Schoups, and N. van de Giesen (2010b), Why hydrological predictions should be evaluated using information theory, *Hydrol. Earth Syst. Sci.*, *14*(12), 2545–2558, doi:10.5194/hess-14-2545-2010.

Zoubir, A. M., and B. Boashash (1998), The bootstrap and its application in signal processing, *IEEE Signal Process. Mag.*, *15*(1), 56–76, doi:10.1109/79.647043.