On the distributions of seasonal river flows: Lognormal or power law?

M. C. Bowers,[1] W. W. Tung,[1] and J. B. Gao[2,3]

Received 23 August 2011; revised 3 April 2012; accepted 8 April 2012; published 19 May 2012.

[1] Distributional analysis of river discharge time series is an important task in many areas of hydrological engineering, including optimal design of water storage and drainage networks, management of extreme events, risk assessment for water supply, and environmental flow management, among many others. Having diverging moments, heavy-tailed power law distributions have attracted widespread attention, especially for the modeling of the likelihood of extreme events such as floods and droughts. However, straightforward distributional analysis does not connect well with the complicated dynamics of river flows, including fractal and multifractal behavior, chaos-like dynamics, and seasonality. To better reflect river flow dynamics, we propose to carry out distributional analysis of river flow time series according to three "flow seasons": dry, wet, and transitional. We present a concrete statistical procedure to partition river flow data into three such seasons and fit data in these seasons using two types of distributions, power law and lognormal. The latter distribution is a salient property of the cascade multiplicative multifractal model, which is among the best models for turbulence and rainfall. We show that while both power law and lognormal distributions are relevant to dry seasons, river flow data in wet seasons are typically better fitted by lognormal distributions than by power law distributions.

Citation: Bowers, M. C., W. W. Tung, and J. B. Gao (2012), On the distributions of seasonal river flows: Lognormal or power law?, *Water Resour. Res.*, *48*, W05536, doi:10.1029/2011WR011308.

1. Introduction

[2] Distributional analysis of river flow time series is a basic task in hydrology [e.g., *Smakhtin*, 2001; *Morrison and Smith*, 2002; *Laio*, 2004; *Carreau et al.*, 2009; *De Domenico and Latora*, 2011; *Sharwar et al.*, 2011]. For example, under the assumption that statistical features of the flow of a river remain fairly constant, when one wants to estimate the probability of drought and develop suitable means to mitigate the effects of drought, one needs to best summarize known knowledge to determine a threshold value and calculate the probability that the river discharge data do not exceed the chosen threshold value [e.g., *Smakhtin*, 2001; *Kroll and Vogel*, 2002; *Benyahya et al.*, 2009]. On the other hand, if one wishes to estimate the probability of floods and assess various kinds of risks associated with them, then one needs to determine another threshold value and calculate the probability that the river discharge data exceed the chosen threshold [e.g., *Fernández and Salas*, 1999; *Fernandes et al.*, 2010; *Um et al.*, 2010]. The former probability is called cumulative distribution function (CDF), while the latter is termed complimentary CDF (or CCDF), which is often called the survival function in statistics and "flow duration curve" (FDC) in hydrology [*Smakhtin et al.*, 1997; *Castellarin et al.*, 2004; *Iacobellis*, 2008]. Distributional analysis is usually carried out using the entire available record, called "period of record" [*Vogel and Fennessey*, 1994]. It can be refined by grouping observations from all similar calendar months from the entire record prior to FDC construction, e.g., all January observations, termed "long-term average monthly" FDCs [*Smakhtin*, 2001].

[3] Phenomenologically, river flow time series often exhibit spikes that rise far above the typical values in the series [*Anderson and Meershaert*, 1998; *Katz et al.*, 2002; *Bernardara et al.*, 2008; *Villarini et al.*, 2011]. In order to adequately capture this behavior in the distributional analysis, distributions which allocate sufficient probability to the upper tail of the river flows must be employed. One such class of distributions is the so-called heavy-tailed distributions, which are distributions with some infinite moments [e.g., *Anderson and Meershaert*, 1998]. In this article we consider a distribution heavy tailed if any of its statistical moments diverge. Since we are examining river flows, which are nonnegative, we only consider one-sided heavy-tailed distributions.

[4] One such heavy-tailed distribution that has gained greatest attention in the scientific literature is the power law distribution [*Newman*, 2005]. It has been used to characterize the population of U.S. cities, the degree distribution of metabolites in the metabolic network of the bacterium *E. coli*, the frequency of occurrence of unique

[1] Department of Earth and Atmospheric Sciences, Purdue University, West Lafayette, Indiana, USA.

[2] Department of Mechanical and Materials Engineering, Wright State University, Dayton, Ohio, USA.

[3] PMB Intelligence LLC, West Lafayette, Indiana, USA.

Corresponding author: M. C. Bowers, Department of Earth and Atmospheric Sciences, Purdue University, 550 Stadium Mall Dr., West Lafayette, IN 47907, USA. (mcbowers@purdue.edu)

0043-1397/12/2011WR011308

words in Herman Melville's novel *Moby Dick*, the number of long-distance calls received by AT&T customers, the severity of terrorist attacks, the size of electrical blackouts, and the sales volume of best-selling books [e.g., *Clauset et al.*, 2009]. It is thus no surprise that the power law distribution is being used extensively in hydrological applications as well [e.g., *Anderson and Meershaert*, 1998; *Pandey et al.*, 1998; *Elshorbagy et al.*, 2002; *Katz et al.*, 2002; *Aban et al.*, 2006; *Bernardara et al.*, 2008].

[5] River flows, while driven by random events such as rainfalls, also have their nonlinear deterministic aspects [*Bernardara et al.*, 2008; *Tung et al.*, 2011]. Indeed, it has been shown that river flows exhibit complicated dynamics, such as fractal and multifractal behavior [*Zhang et al.*, 2008], chaos-like dynamics [*Sivakumar*, 2004; *Wang et al.*, 2006; *Tung et al.*, 2011], and periodicity in the mean, standard deviation, and skewness [*Tesfaye et al.*, 2006]. Traditional distributional analyses of river flow data do not connect well with such complicated river flow dynamics. Thus, we propose that a more refined implementation of the well-known distributional methods may enhance the information these methods offer about river flow dynamics.

[6] In attempting to refine the standard distributional techniques, we first look to the most ubiquitous features of river flows. One of the most documented properties of river flow data is the annual cycle [e.g., *Katz et al.*, 2002; *Tesfaye et al.*, 2006]. This can be readily illustrated by the sharp spectral peaks corresponding to 1 year cycle and its harmonics (an example will be discussed in section 3). We can take advantage of this cyclic behavior by using it to identify and capture distinct flow regimes. Intuitively, the behavior of river flows should differ most between the peak and the trough of the annual cycle. One can imagine that the peak of the annual cycle corresponds to a more intense flow regime or "wet season", while the trough corresponds to a more mild flow regime or "dry season." The time spent in between the wet season and the dry season can be considered as the transition season. To better reflect river flow dynamics in our distributional analyses, we propose that the calendar (or hydrologic) year be divided into three "flow seasons", and then perform distributional analysis separately for each regime.

[7] Considering separate distributional analysis for each flow regime, rather than lumping all data together, allows us to preserve information about annual variability that would otherwise be lost. This partition is achieved by deriving a smooth function from the daily climatological mean and variance of the flow levels, and then checking whether the smooth function exceeds or falls below certain suitably chosen threshold values. The major difference between our method and the long-term average monthly FDCs is that the timing in our method is identified from flow regimes, while the latter method simply uses the timing of the calendar months.

[8] In the present paper, we focus on two particular distributions to describe the constructed seasonal river flows: power law and lognormal. The latter is a salient property of the cascade multiplicative multifractal model, which is among the best models for turbulence and rainfall [*Gao et al.*, 2007]. Specifically, we shall analyze nine river flow time series collected from sites in the continental United States. We show that, in our data, while both power law and lognormal distributions are very relevant to dry seasons, data in wet seasons are typically well fitted by lognormal distributions.

[9] This paper is organized as follows. In section 2 we describe the river flow data used in this study, present a method for dividing a given river flow time series into three flow seasons, and describe procedures for seasonal river flow distributional analysis. In section 3 we present the results of the distributional analysis and Kolmogorov-Smirnov testing, and in section 4 we discuss these results to gain a good understanding. General concluding remarks are given in section 5.

2. Data and Methodology

2.1. Data

[10] In this study we analyzed daily mean river discharge observations (in cubic feet per second, ft^3/s) from nine rivers across the continental United States. For the geographical locations of the river flow measurement sites, see Figure 1 and Table 1. The data set for the Mississippi River near St. Louis is the longest, extending from 1861 to 2010.

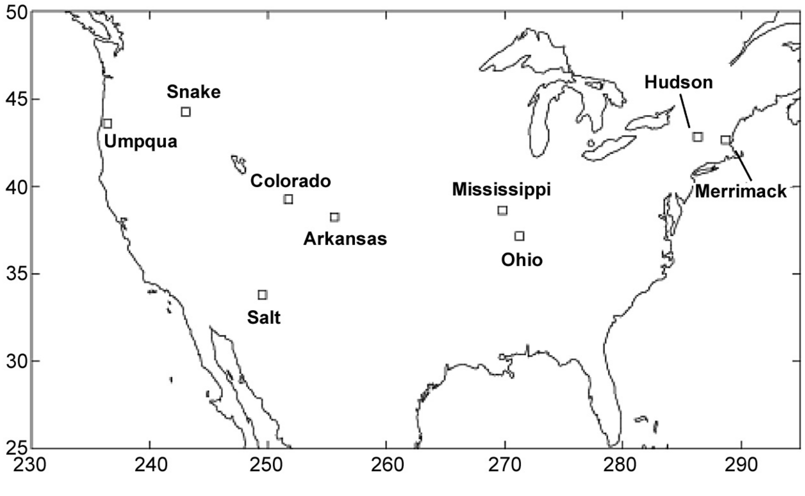

Figure 1. Map showing locations of river flow measurement sites.

Table 1. Select Characteristics of River Measurement Sites Studied

River	Station	Latitude	Longitude	Period of Record
Arkansas	07109500	38°14′53″	104°23′55″	1931–1951, 1965–2010
Colorado	09095500	39°14′21″	108°15′56″	1933–2010
Hudson	01335754	42°49′55″	73°40′00″	1887–1956, 1976–2010
Merrimack	01100000	42°38′45″	71°17′56″	1923–2010
Mississippi	07010000	38°37′44″	90°10′47″	1861–2010
Ohio	03611500	37°08′51″	88°44′27″	1928–2010
Salt	09497500	33°47′53″	110°29′57″	1924–2010
Snake	13269000	44°14′44″	116°58′51″	1910–2010
Umpqua	14321000	43°35′10″	123°33′15″	1905–2010

The data sets were obtained from the USGS surface water information site (http://water.usgs.gov/osw/). Prior to further processing, the raw time series data were checked for possible long-term trends (which amounts to nonstationarity) in daily streamflow; detrending analysis found no systematic changes contributing to either linear or quadratic trends. Additional experiments were performed to evaluate the results' sensitivity to the El Niño–Southern Oscillation. The entire distributional analysis was repeated using only data observed during El Niño years and again for only non El Niño years. The difference was negligible, and therefore, the effects of El Niño years were not considered separately.

[11] Note that the raw river flow data may include variability not only from natural sources, but also from anthropogenic ones [*Poff et al.*, 1997]. For example, dams, reservoirs, and pumping systems are operated on some of the rivers. The extent of such modifications on the natural river flow regime can, of course, range from negligible to profound [*Black et al.*, 2005]. For the present paper, our concern is to avoid data including sudden and dramatic river flow regime changes induced primarily by hydrologic engineering projects. Evidence of this kind of bifurcation can be detected in raw river flow time series [e.g., *Vörösmarty and Sahagian*, 2000; *Li et al.*, 2007; *Villarini et al.*, 2011], and we avoid using data with such features. This leaves only variation induced by those anthropogenic modifications which do not cause a sudden and dramatic change in river flow behavior. We treat these features in the same way we treat natural modifications like erosion and deposition: they are all contributors to the overall variability of the time series. The effect of anthropogenic hydrologic modifications on our analysis is considered further in section 4.1.

2.2. Defining a River's Flow Seasons: Wet, Dry, and Transitional

[12] Intuitively, a wet and a dry season may be associated with large and small values of river discharges. In between would be the transitional seasons, either from wet to dry, or vice versa. The idea can be conveniently implemented using annual climatology.

2.2.1. Annual Climatology

[13] Consider a river flow time series $x(t)$, $t = 1, 2, \ldots, n$. Each point $x(t)$ is associated with the particular calendar day $i = 1, 2, \ldots, 365$ when it was observed. Data on the same day of different years are grouped together and their mean and standard deviations (STDs) calculated. For simplicity, the 29th day of February in the leap years may be discarded. We call the variation of such mean and STD on the days of a year the annual climatology of the flow series. We denote them as $\bar{x}_i, s_i$, $i = 1, \ldots, 365$, respectively.

2.2.2. Identifying Flow Seasons by Thresholding the Criterion Function

[14] Now that we have introduced climatological mean and variance for each calendar day of the year, we may associate wet and dry seasons with relatively high and low values of climatological mean and variance. More concretely, we may define a new function, φ_i, by combining the mean $\bar{x}_i$ and STD s_i.

$$\varphi_i = \bar{x}_i + 2s_i \tag{1}$$

A factor of 2 is included before s_i to enhance the seasonal variability in φ_i. Since $\bar{x}_i$ and s_i are sample means and standard deviations, they are random variables. Therefore, φ_i may be quite irregular because of the potentially large day to day variability in either $\bar{x}_i$ or s_i. To ease subsequent analysis, we smooth φ_i using the adaptive detrending algorithm [*Hu et al.*, 2009a; *Gao et al.*, 2010; *Tung et al.*, 2011; *Gao et al.*, 2011] with second-order local polynomials and window size of 129 days to obtain a new globally smooth function, c_i, which may be called the criterion function:

$$\text{Filtering } \varphi_i \to c_i \tag{2}$$

The adaptive algorithm employed here has been found to be able to reduce noise more effectively than linear filters, wavelet- and chaos-based noise reduction techniques, perform multiscale decomposition, and yield excellent trends from data sets with arbitrary trends. On the basis of c_i, we can define wet and dry seasons as follows:

$$I_{\text{wet}} = \{i : c_i > D_{\text{wet}}\} \tag{3}$$

$$I_{\text{dry}} = \{i : c_i < D_{\text{dry}}\} \tag{4}$$

where I_{wet} and I_{dry} are the largest contiguous subsets of i satisfying the right-hand sides of equations (3) and (4), respectively, and D_{wet} and D_{dry} are threshold values. Other time indices would form the transitional season. Note that $i = 365$ and $i = 1$ are considered contiguous points. Now a critical question is how to choose D_{wet} and D_{dry}.

[15] Note that for the wet and dry seasons to be defined robustly, the "valid" values of D_{wet} and D_{dry} have to form some interval, instead of just few single points. This means that if we have chosen a value for D_{wet} and subsequently determined a wet season, the distribution for the wet season flow data has to remain fairly robust if we slightly increase or decrease D_{wet} to determine another wet season. In the following analysis, we have chosen the smallest such D_{wet} to define a wet season. Similarly, we have chosen the largest such D_{dry} to determine a dry season. This strategy makes the transitional season "minimal", so that its distribution can be ignored. In fact, when the distributions for the wet and dry seasons are different, it may not be

meaningful to consider the distribution of the transitional season. Note that initial values of D_{wet} and D_{dry} may be chosen to be near the maximum and minimum of the criterion function c_i. The optimal D_{wet} and D_{dry} can be determined by systematically decreasing D_{wet} and increasing D_{dry} and checking when the distributions for the wet and dry seasons change quantitatively.

2.3. Distributional Analysis

[16] After identifying the annual timing of a river flow's wet and dry seasons we define two sets of random variables, denoted by X_{wet} and X_{dry}, which contain daily river discharges that occurred during the identified wet and dry seasons, respectively. We then carry out analysis of their distributions separately. The sample sizes of these new seasonal flow populations, x_{wet} and x_{dry}, varied across rivers and seasons, depending on the original length of the time series and the number of calendar days included in each season. The sample sizes were generally on the order of 10^4, with a minimum of 3619 for the Colorado River wet season, a median of 10,429, and a maximum of 26,826 for the Hudson River dry season. We found that seasonal data were best described by either of two probability distributions: the lognormal distribution or the power law distribution. In either case, we used graphical comparisons between the empirical and theoretical probability functions to guide model selection, methods of maximum likelihood for parameter estimation, and Kolmogorov-Smirnov testing to evaluate goodness of fit of the selected distributional models. Empirical densities were constructed following *Vogel and Fennessey* [1994] by binning the log-transformed data and linearly rescaling.

[17] We now define our parameterization of the lognormal and the power law random variables. A random variable X follows a lognormal distribution if its logarithm follows a normal distribution. While it is most convenient to use natural logarithm, to ease discussion of the figures presented below, we use logarithm base 10. Let Y be a normally distributed random variable with mean μ and standard deviation σ, then $X = 10^Y$ has a lognormal distribution with its density given by

$$f_X(x(t)|\mu,\sigma) = \frac{1}{x(t)\ln(10)\sqrt{2\pi}\sigma}\exp\left[-\frac{(\log_{10}x(t)-\mu)^2}{2\sigma^2}\right], x(t) > 0 \tag{5}$$

[18] It is interesting to note that lognormality is a salient feature of the multiplicative cascade multifractal [*Gao and Rubin*, 2001a, 2001b; *Tung et al.*, 2004], which is considered one of the best models for fully developed turbulence and rainfall [*Gao et al.*, 2007]. The basic model of the cascade process prescribes that the process may be expressed as

$$x = r_1 r_2 \cdots r_n, \tag{6}$$

where $0 \leq r_i \leq 1,\ i = 1, \ldots, n$ are called multipliers having the same probability density function. Taking logarithm of both sides of equation (6) and using the central limit theorem, one immediately sees that $\log(x)$ converges to a normal distribution, thus x follows a lognormal distribution.

[19] To check for possible lognormal behavior, we first calculated parameter estimates $\hat{\mu}$ and $\hat{\sigma}$ (where "hats" indicate that quantities are estimates of unknown parameters) using maximum likelihood estimation like *Stedinger* [1980]. A lognormal density with μ and σ set equal to $\hat{\mu}$ and $\hat{\sigma}$, respectively, was then plotted along with the empirical density for comparison. Kolmogorov-Smirnov testing was then employed to evaluate goodness of fit of the selected lognormal model.

[20] A random variable X follows a power law distribution if the probability of exceeding a value $x(t)$ is proportional to a constant power of $x(t)$. The CCDF of X is denoted as

$$P[X > x(t)] = \left(\frac{x(t)}{b}\right)^{-\alpha}, \qquad 0 \leq b < x(t) \tag{7}$$

where α is the shape parameter and b marks a lower bound on the possible values that a Pareto distributed random variable can assume. Like the lognormal distribution, the power law CCDF has two parameters. Equation (7) only describes values of river flow greater than b, and does not take into account flows less than b. In fact, If one wishes to model these values as well, additional parameters must be introduced. Note that for a given scaling parameter α, all statistical moments, μ_k with $k \geq \alpha$, are infinite. That is why such power law distributions are heavy tailed. In particular, when $0 < \alpha < 2$, the variance of X is infinite; when $0 < \alpha \leq 1$, the mean of X is also infinite.

[21] To assess if data exhibit power law tails, we plotted the log-transformed empirical survival function, $\log_{10}[P(X > x(t))]$, against $\log_{10}(x(t))$ and examined whether or not there exists a linear scaling relation [e.g., *Mandelbrot*, 1963; *Pandey et al.*, 1998; *Aban et al.*, 2006]. Once a possible power law relationship was identified, we used the maximum likelihood method presented by *Clauset et al.* [2009] to estimate b and α. To evaluate goodness of fit, we used Kolmogorov-Smirnov (K-S) testing as done by *Goldstein et al.* [2004].

3. Results

3.1. Seasonal River Flow Distributions

[22] Figure 2 shows mean daily river discharge for the nine rivers on the longest timescale represented in the study (1861–2010). Note the occasional spikes which rise far above the typical values in the series; this intermittent behavior can be seen, to varying degrees, in all nine of the river time series used in the study. As expected, the river flow data possess strong annual cycles. An example is shown in Figure 3. Figure 4 shows the annual climatology of the nine rivers. The criterion functions have been thresholded as described in section 2.2.2 to obtain the timing of the flow seasons, unique to each river. In keeping with our selection criteria, the observations included in the wet seasons have higher sample mean and variance than those included in the dry seasons.

[23] Each of the nine rivers' seasonal flow distributions followed one of four patterns: (1) lognormal wet and dry seasons, (2) power law wet and dry seasons, (3) lognormal wet season and power law dry season, and (4) lognormal dry season and unidentified wet season distribution. By unidentified, we mean that neither lognormal, power law,

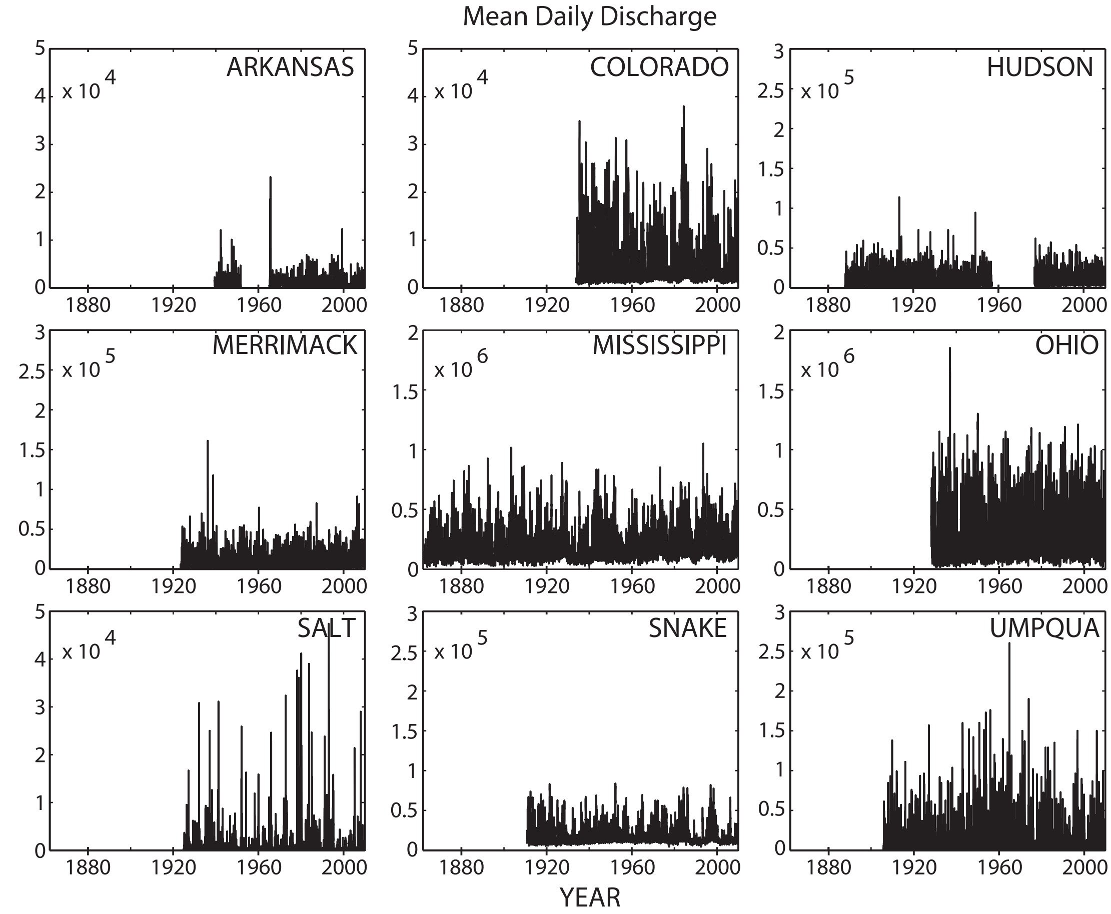

Figure 2. Mean daily discharge time series in ft^3 s^{-1} for the nine rivers studied. Note that the typical flow values for the nine rivers vary over 2 orders of magnitude. For visibility, three different scales are used across the plots.

nor Weibull models fit the data satisfactorily according to our graphical methods. These results are summarized in Table 2. Plots illustrating the first three scenarios are given in Figures 5, 6, and 7.

3.2. K-S Testing and River Flow Data

[24] Published research on distributional analysis of river flows has emphasized parameter estimation but has rarely examined the goodness of fit of the chosen distributions. In this work we examine the latter using the K-S test. We have found that even though lognormal distributions tend to fit river flow data better than power law distributions in the wet season, the p value from the K-S test is generally smaller than 0.05, signifying that the lognormal fitting fails the K-S test. The test for dry season fitting using power law is only slightly better. Therefore, a fundamental question is: why would distributional analysis of streamflow data often fail the K-S test?

[25] While there may be multiple mechanisms for the negativeness of the K-S test, we believe the main mechanism has to be the multiplicity of the physical processes that can affect recorded river discharges. We emphasize here that the variations in river flows resulting from different physical processes may assume different distributions. This means, in general, the distribution for river discharge is actually a mixture of several different distributions. It is known that the K-S test is sensitive to noise and measurement effects [*Lampariello*, 2000]; thus, we wish to examine how noise may affect the skill of the K-S test. For ease of exposition, let us consider random variables X, Z, and N, whose time series innovations $x(t)$, $z(t)$, and $n(t)$ satisfy the following relationship:

$$z(t) = x(t) + n(t)\,, \tag{8}$$

where $x(t)$ describes the main variation of a river discharge, $n(t)$ may be considered measurement noise (e.g., Gaussian noise) or variation caused by a different physical process than that which caused $x(t)$, and $z(t)$ is the measured data. Pertinent to the theme of the paper, $x(t)$ may be assumed to follow a lognormal or power law distribution, and our task is to verify whether $z(t)$ follows the same distribution as that of $x(t)$. The essence of the K-S test is to examine how close the CDF $F_Z(z)$ of $z(t)$ is to the CDF $F_X(x)$ of $x(t)$:

$$\delta = \sup_{-\infty<a<\infty} |F_Z(a) - F_X(a)| \tag{9}$$

where δ is called the K-S statistic, and as $\delta \rightarrow 0$, $x(t)$ and $z(t)$ are considered to follow the same distribution [*DeGroot and Schervish*, 2002; *Goldstein et al.*, 2004; *Zhou et al.*, 2004; *Hu et al.*, 2009b].

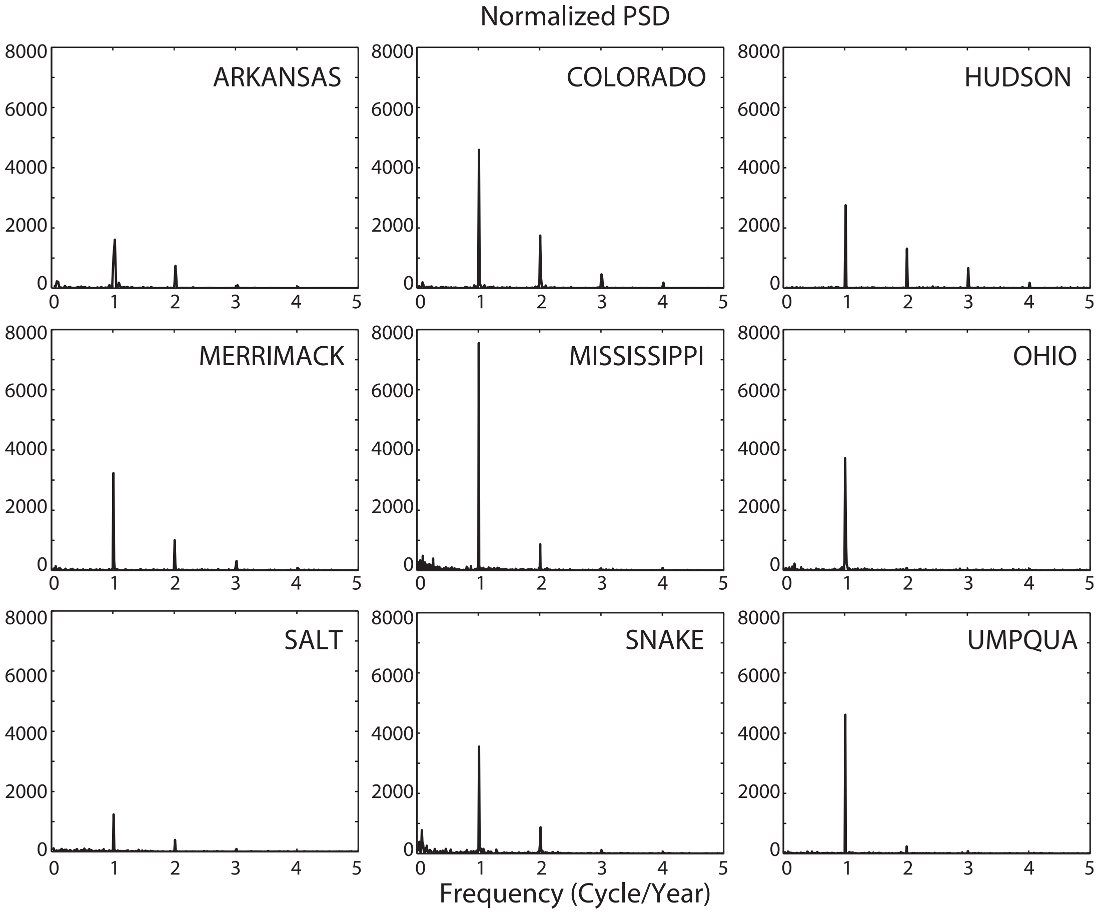

Figure 3. Power spectral densities for the nine rivers. Note the spikes at 1 cycle per year, indicative of strong annual cycles. Each spectrum has been normalized by its mean power.

[26] To better appreciate equation (9), let us consider a situation where $Z = X$ with probability $1 - q$ and $Z = C$, where C is a constant, with probability q. Then, $\delta = q$. Casual inspection can readily verify that a $q \approx 0.05$ will already likely reject the null hypothesis in the sense that the p value is less than significance level 0.05 (note that the functional relation between δ, which is equal to q here, and the p value is highly nonlinear).

[27] We used Monte Carlo simulations to study the effect of modulating a lognormal or power law $x(t)$ by a noise signal $n(t)$. The task is to determine how large the noise signal $n(t)$ must be in order to cause $z(t)$ to systematically fail the K-S test. It is instructive to use the ratio of the standard deviation of the noise $n(t)$ to the sample standard deviation of the measured signal $z(t)$ to appreciate the relative magnitude of the noise necessary to undermine the K-S test. Let

$$\epsilon = \frac{\mathrm{STD}[n(t)]}{\mathrm{STD}[z(t)]} \tag{10}$$

where STD( ) is the standard deviation operator.

[28] Using a power law $x(t)$, and Gaussian $n(t)$, we found that when the ratio ϵ exceeds about 0.02, the average p value of the test falls below significance level 0.05 and $z(t)$ fails the K-S test more often than not. This signifies that for a river flow signal following a power law distribution, modulation by another process, measurement error or otherwise, with a standard deviation around 2% that of the measured process will likely cause the K-S test to reject the power law hypothesis. We used similar simulations to study the case when $x(t)$ follows a lognormal distribution and found that an ϵ of only about 0.06 is needed for the average p value to fall below 0.05 and for the K-S test to reject the lognormal hypothesis more often than not. Table 3 displays the results of these Monte Carlo simulations. From these simple experiments, we see that even a small amount of noise is enough to prevent the K-S test from correctly identifying the true underlying distribution of a signal. Since river flow signals can be quite noisy, we can thus consider the results of the robust graphical methods described in section 3.1 not only valid, but indispensable. Of course, one can be even more confident about a given model if the K-S test yields a small p value.

4. Discussion

[29] Many physical processes can contribute to measured river flows, including rainfall, snowmelt, evaporation, and permeation, as shown in Figure 8. As one can readily envisage, the contribution from a specific physical process to the observed discharge varies in time, space, and from river to river. This is the very reason that observed discharge

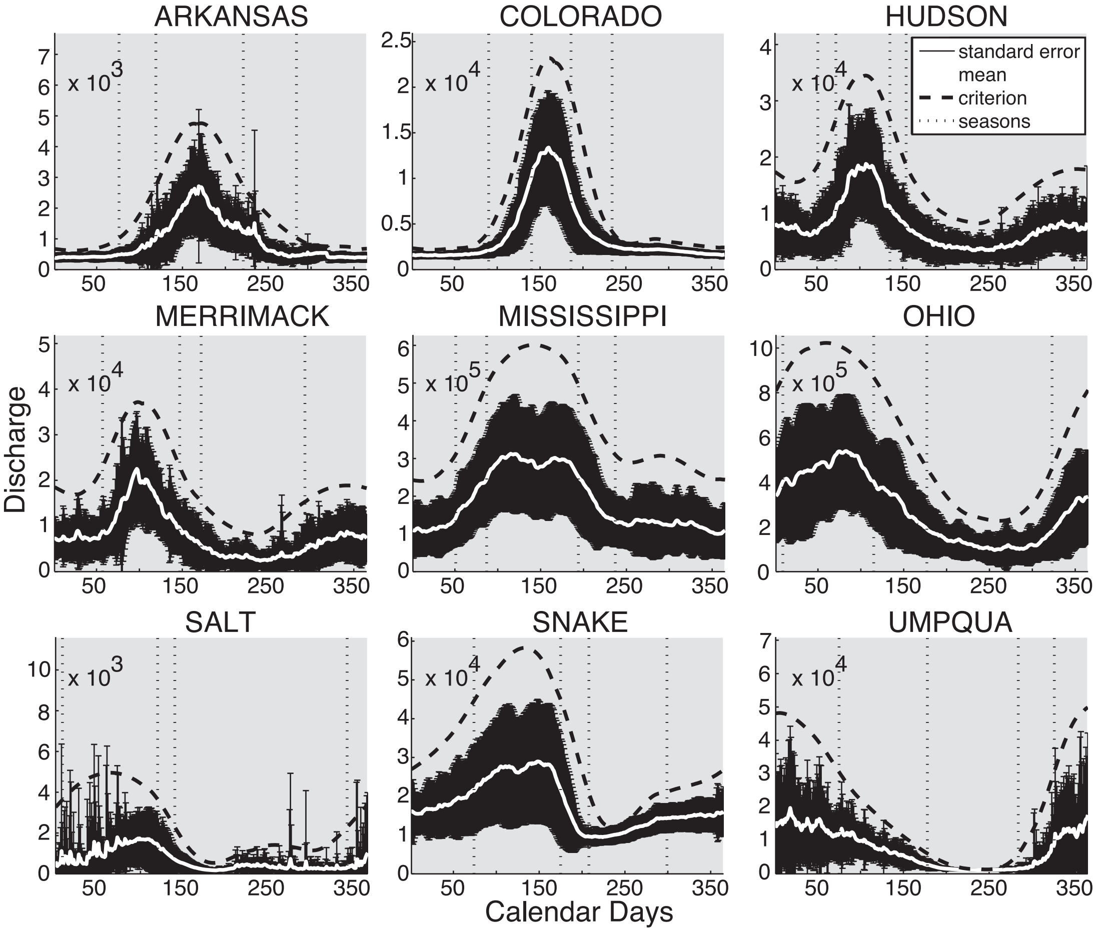

Figure 4. Annual climatology composed of daily mean $\bar{x}_i$ (white line) and daily standard deviation s_i (black error bars) with the criterion function c_i (dashed line). The criterion function is thresholded such that calendar days on which c_i exceeds D_{wet} belong to the wet season I_{wet} and calendar days for which c_i falls below D_{dry} belong to the dry season I_{dry}. Vertical dashed lines indicate divisions between the seasons.

distributions can vary from season to season and from river to river. From discharge data alone, it is impossible to deduce the exact effect of each of the contributing physical processes. Our aim in this section is to gain some insights into the effects of two of the most important processes–rainfall and snowmelt. We also explore how these effects may be modulated by anthropogenic modifications.

4.1. Physical Interpretations

[30] As shown in Table 2, all nine of the rivers studied here exhibit different statistical properties in their wet seasons versus their dry seasons. One distinction is that some of the rivers, i.e., Mississippi, Merrimack, Hudson, and Salt, show a difference in only the parameters of their seasonal distributions (which may be called variability of the

Table 2. Results of Seasonal Distributional Analysis

	Wet Season			Dry Season		
River	Behavior	Parameter	Estimates	Behavior	Parameter	Estimates
Arkansas	lognormal	$\hat{\mu} = 3.1582$	$\hat{\sigma} = 0.3033$	power law	$\hat{\alpha} = 2.50$	$\hat{b} = 360$
Colorado	unidentified			lognormal	$\hat{\mu} = 3.2589$	$\hat{\sigma} = 0.1242$
Hudson	lognormal	$\hat{\mu} = 4.1153$	$\hat{\sigma} = 0.2593$	lognormal	$\hat{\mu} = 3.6652$	$\hat{\sigma} = 0.2876$
Merrimack	lognormal	$\hat{\mu} = 4.0900$	$\hat{\sigma} = 0.2660$	lognormal	$\hat{\mu} = 3.4150$	$\hat{\sigma} = 0.3202$
Mississippi	lognormal	$\hat{\mu} = 5.4125$	$\hat{\sigma} = 0.2116$	lognormal	$\hat{\mu} = 5.0216$	$\hat{\sigma} = 0.2272$
Ohio	unidentified			lognormal	$\hat{\mu} = 5.0262$	$\hat{\sigma} = 0.2666$
Salt	power law	$\hat{\alpha} = 2.37$	$\hat{b} = 2840$	power law	$\hat{\alpha} = 1.97$	$\hat{b} = 725$
Snake	unidentified			lognormal	$\hat{\mu} = 4.0422$	$\hat{\sigma} = 0.1198$
Umpqua	lognormal	$\hat{\mu} = 3.9786$	$\hat{\sigma} = 0.3816$	power law	$\hat{\alpha} = 2.50$	$\hat{b} = 1990$

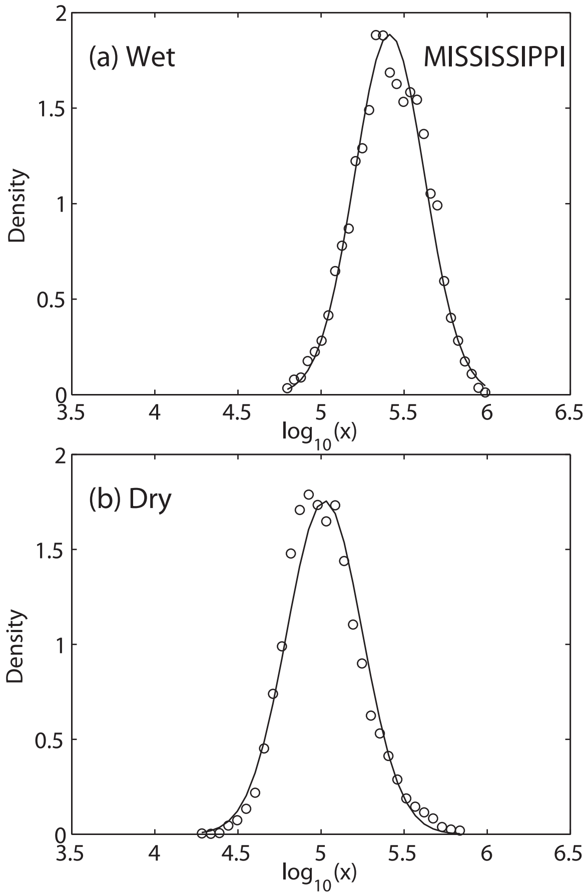

Figure 5. Empirical density (circles) and lognormal probability density function (solid line) with $\mu = \hat{\mu}$ and $\sigma = \hat{\sigma}$ for Mississippi River (a) wet and (b) dry flow season data.

first kind), while others, i.e., Umpqua and Arkansas, exhibit seasonal flow deviates drawn from completely different families of distributions (which may be called variability of the second kind).

[31] A possible explanation of why some rivers' seasonal distributions come from different families while others come from the same family lies in the temporal variability of the physical mechanisms governing stream flows. Although these mechanisms vary on all time scales, seasonal variability is particularly prevalent. If the nature of the most influential physical processes during one part of the calendar year is sufficiently different from that of another part of the year, we could expect the difference to manifest in the statistical properties of the seasonal distributions by producing variability of either the first or second kind. Following are two conjectured scenarios.

[32] One way to realize the statistical variability of the first kind could be to adjust only the frequency and/or severity of the most prominent physical mechanisms generating flow values. Consider a river basin which collects the vast majority of its water from local precipitation events. Imagine that these precipitation events tend to have a certain frequency and severity during the river's dry season and another (greater) frequency and severity during the river's wet season. This rescaling of the inputs could translate to a rescaling of the outputs, i.e., the seasonal river flows may follow the same family of distributions but with rescaled parameters.

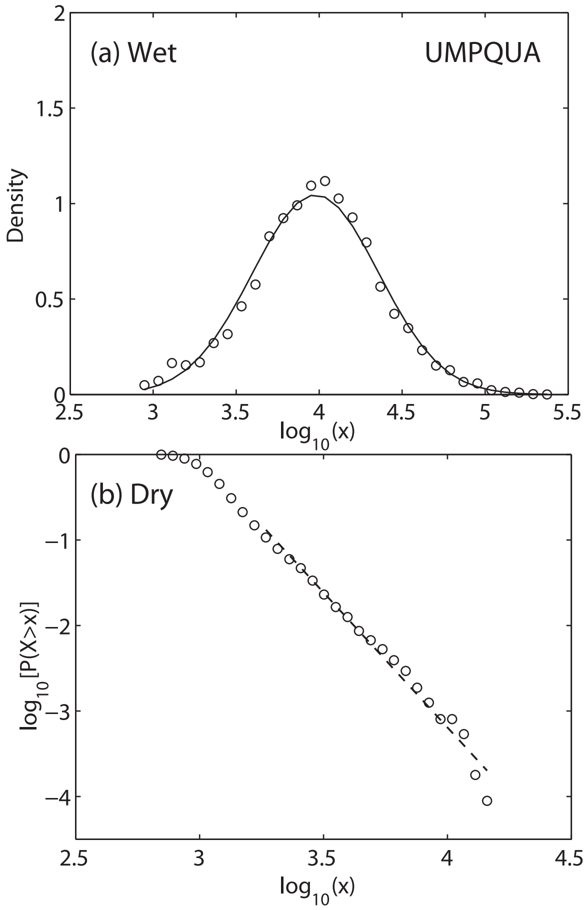

Figure 6. (a) Empirical density (circles) and lognormal probability density function (solid line) with $\mu = \hat{\mu}$ and $\sigma = \hat{\sigma}$ for Umpqua River wet flow season data. (b) $\log_{10}[P(X > x(t))]$ versus $\log_{10}X$ for Umpqua River dry flow season data.

[33] To achieve the second kind of statistical variability, in which the families of the seasonal distributions differ, we could envision a change of the fundamental nature of the predominant physical processes. Consider a river that experiences a precipitation regime with constant frequency and severity throughout the year. Imagine that the river lies in the drainage basin for an area that collects snowpack throughout the winter. Every spring, temperatures rise and a large volume of water is released from the snowpack and drains into the river. This snowmelt process dominates the river's flow regime during the spring and constitutes its wet season. Since the fundamental nature of the processes delivering water to the river differ seasonally (snowmelt in the wet season and background precipitation events during the dry season), we may expect the river flow output to exhibit completely different statistical behavior seasonally, i.e., the seasonal observations appear to be drawn from different families of distributions.

[34] We have noted the prevalence of annual variability in the physical mechanisms generating river flows. The seasonality of most natural phenomena can be traced back, at least indirectly, to the annual cycle in solar radiation

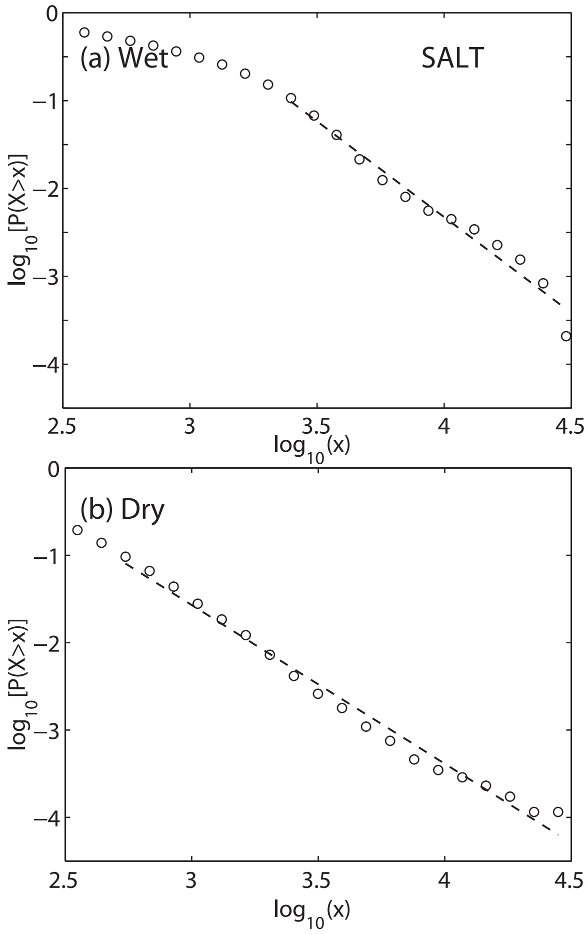

Figure 7. $\log_{10}[P(X > x(t))]$ versus $\log_{10}X$ for Salt River (a) wet and (b) dry flow season data.

incident at the top of the atmosphere, which tends to be smooth and sinusoidal (excepting the polar regions). We can therefore expect to observe the thermodynamic state variables, such as temperature, within the Earth systems to inherit some of this primarily smooth and sinusoidal character. Similarly, water goes through phase changes under the direct or indirect influence of solar energy, resulting in changes of thermodynamic energy. This is the fundamental linkage between the global hydrologic and energy cycles, as well as the local water and thermodynamic energy balances. It is therefore no surprise that the frequency and severity of some area's precipitation events vary smoothly through the year like the temperature. On the other hand, a large number of natural mechanisms operate in more or less discrete, e.g., "on or off," states, determined by some smoothly changing input signal. Seasonal snow melt, for example, tends to occur only when temperatures are in a given range after some duration, and does not contribute to river flows when temperatures fall out of that range.

Table 3. Monte Carlo Performance of Kolmogorov-Smirnov Test on Signal Z Composed of Power Law or Lognormal X Modulated by Gaussian Noise N

Power Law X			Lognormal X		
ϵ	p Value	Rejection Rate	ϵ	p Value	Rejection Rate
0.0238	0.0464	65.6%	0.0613	0.0471	62.1%
0.0254	0.0312	79.9%	0.0630	0.0358	75.3%
0.0270	0.0206	91.0%	0.0647	0.0268	85.7%
0.0285	0.0132	95.9%	0.0664	0.0197	94.3%
0.0301	0.0084	99.0%	0.0681	0.0143	98.2%

[35] For many rivers, especially those most closely tied to human livelihood, mechanisms linked to anthropogenic effects may also contribute to the variation observed in river flows. Major effects can be caused by reservoir construction and operation, groundwater abstraction and storage, surface water abstractions, discharges and transfers, and from spatially extensive land use changes such as deforestation or agricultural intensification [*Black et al.*, 2005]. Here again, the characteristics of the variation depend on the particular mechanism in question. For example, it is known that drying due to overgrazing and other intensive land use on semiarid soils leads to a net loss of soil water storage, reduced evaporation, and increased storm runoff [*Vörösmarty and Sahagian*, 2000]. This increases a river's response to precipitation events; thus, for a river whose flow is controlled mostly by precipitation, the effect of the land use change manifests more prominently during the river's wet flow regime. This seasonality in the intensity of the anthropogenic effect can further the distinction between the river's wet and dry flow regimes, imparting said variability of either the first or second kind.

[36] Other anthropogenic mechanisms may actually inherit seasonality from their natural counterparts. Take for example a reservoir which follows a schedule of retention during a river's naturally wet regime and release during its naturally dry regime. Although the reservoir's activity is very much seasonal, its direct effects are in opposition to the seasonal tendencies of the river. This opposition has the combined effect of subduing the prevalence of the annual cycle in river flows, as well as distorting the statistical distributions of the seasonal river flows. As we showed in section 3.2, this kind of systematic perturbation of flow values could quite easily prevent a Kolmogorov-Smirnov test from identifying the true underlying distribution of the seasonal river flows.

[37] In this study, we have taken an approach that involves dividing a given river flow time series into repeating annual seasons and fitting a single distribution to the observations drawn during a given season. We are implicitly assuming that seasonal observations do in fact follow, or are at least well approximated by, a single probability distribution. Given that certain mechanisms vary smoothly throughout the year while others vary discretely, it seems reasonable that a distributional analysis with higher temporal resolution could be more accurate and more physically meaningful. One can imagine dividing the calendar year into more and more subperiods prior to the distributional analysis. A stochastic model could be applied to capture the variation in the parameters of the "subseasonal" flow distributions. After optimization, the temporal variation of parameters of this stochastic model could be linked back to physical mechanisms which exhibit similar types of variation. This kind of model could potentially be both more accurate and more physically meaningful than the seasonal analysis presented in this paper.

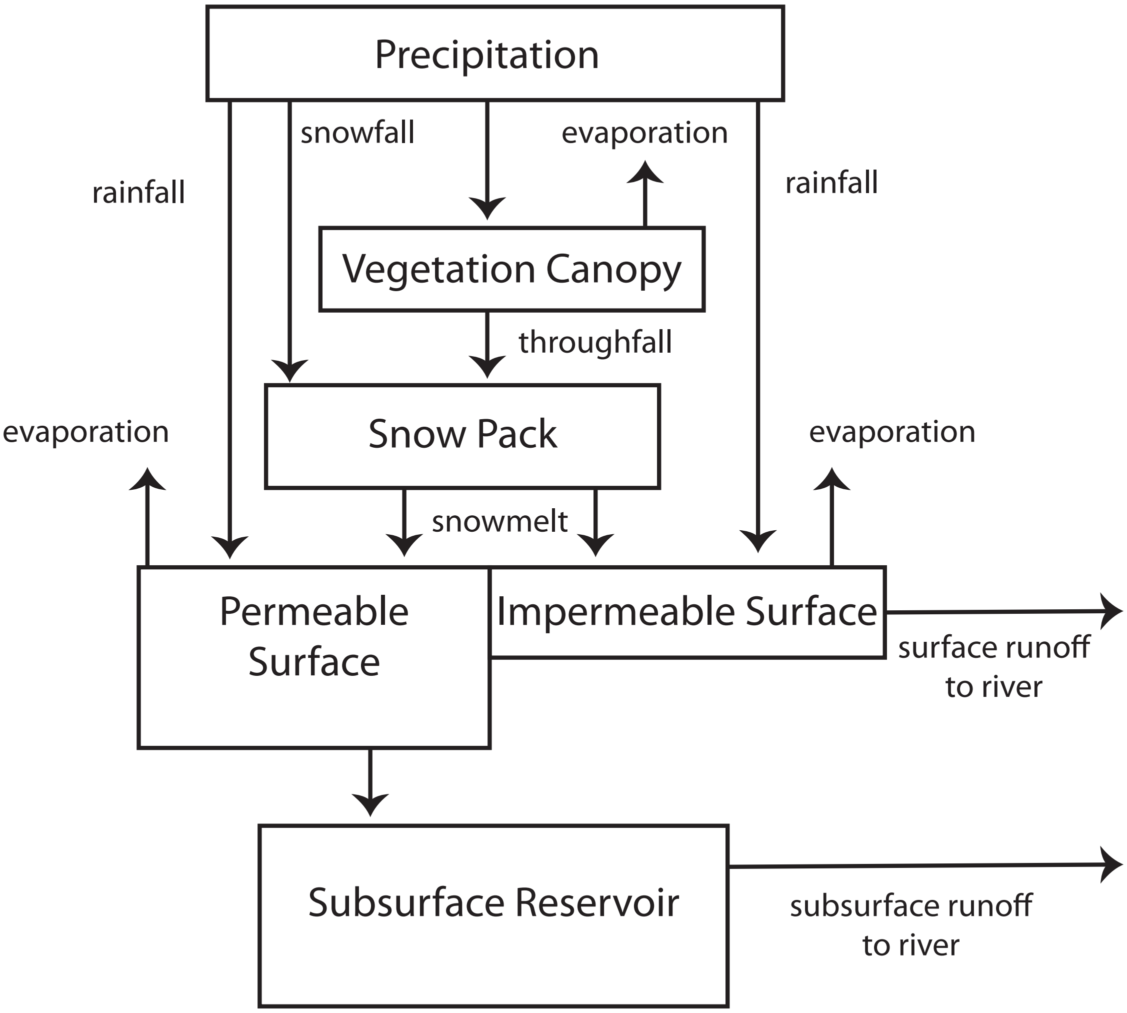

Figure 8. Schematic illustrating the major processes influencing river runoff.

4.2. Heavy-Tailed Behavior

[38] It is interesting that heavy-tailed relationships, particularly the power law, are used so frequently in hydrological applications like flood frequency analysis, yet with the exception of the Salt River, the only power law tails found in the present study were for data observed during the dry season. Furthermore, for the Umpqua and Arkansas rivers, the upper tails of the dry season power law distributions are no more extreme than flow values that would be considered fairly typical during these rivers' corresponding wet seasons. For instance, the maximum value observed during the Umpqua's dry season, 17,000 ft^3 s^{-1}, corresponds to the 0.753 quantile of the wet season observations. Flows of this magnitude would hardly be associated with catastrophic flooding events.

[39] For the exception, the Salt River, heavy-tailed power law relationships were found in both the dry and wet season flow values. This is consistent with the findings of *Anderson and Meershaert* [1998], who reported a power law tail in the distribution of monthly flow values on the Salt River. It is important to note the discrepancy in the parameter estimates of α between *Anderson and Meershaert* [1998] and the present paper. On the basis of monthly river discharges, *Anderson and Meershaert* [1998] estimated the true value of α to be 3.023, which is larger than our estimates of $\hat{\alpha} = 2.37$ during the wet season and $\hat{\alpha} = 1.97$ during the dry season. This discrepancy is caused by the difference in the temporal resolution of the data sets used (e.g., monthly versus daily).

[40] Because of the relative lack of heavy-tailed behavior in the river flows studied here, we present a cautionary note on the use of heavy tails, particularly the power law relationship, in characterizing river flow phenomena. The identification of heavy-tailed behavior should be based on data and analysis. The utmost care should be exercised when conjectures are made concerning heavy-tailed behavior without the direct support of data, as in using regional analysis to estimate flow distributions for ungauged sites.

4.3. Lognormal Behavior

[41] The abundance and prevalence of lognormal behavior in dry and wet seasons suggest that the various kinds of physical mechanisms governing river flows are often self-organized such that equation (6) is at least approximately true.

[42] It should be noted that the moments of a lognormal distribution are all finite, therefore, lognormal distributions

are not heavy-tailed distributions. Nevertheless, for a finite river discharge range, with suitable parameter values, a lognormal distribution may yield approximately a power law distribution. For the details, we refer to *Montroll and Shlesinger* [1982, 1983] and *Newman* [2005].

[43] Although we found the lognormal distribution more relevant to seasonal river flows in our data, we present an additional cautionary note on the justification of its use for this purpose. The lognormal distribution has been used to characterize a number of meteorological variables including height in cumulus cloud populations [e.g., *Lopez*, 1977], atmospheric water droplet size [e.g., *Ochou et al.*, 2007], precipitable water [e.g., *Foster and Bevis*, 2003], and rainfall rate [e.g., *Atlas et al.*, 1990]. To assume that lognormally distributed rainfall, for instance, over a river's catchment area would necessarily produce lognormally distributed river flows is folly. Note that river flows are the realizations of a multiscale nonlinear dynamical system, and lognormal precipitation inputs would certainly not necessitate lognormal river flow outputs. Here again, we stress that selection of a particular distributional model, like lognormal, for river flows should be directly based on data and analysis.

5. Conclusions

[44] To better reflect complicated river flow dynamics, in this work, we attempted to isolate two "extreme" seasons, wet and dry, from continuous measurements of river discharge data, and carried out their distributional analysis separately. We found that discharge data in dry seasons can be well fitted by either power law or lognormal distributions. The discharge data in wet seasons, are however, almost always well fitted by lognormal distribution (with the Salt river being an exception).

[45] Graphical methods were utilized in the initial distributional model selections, while Kolmogorov-Smirnov testing was employed to evaluate goodness of fit of the selected models. We show how relatively subtle noise from other physical processes, including systematic measurement errors, can easily prevent the K-S test from identifying a signal's true underlying distribution. These discussions also connect well with our assertion that river flows are the realizations of a multiplicity of processes, which should in fact yield a mixture of distributions.

[46] Because of the intermittent nature of high-magnitude events in river flow time series and the prevalence of heavy-tail characterization in the hydrologic literature, we paid close attention to the presence or absence of heavy-tailed behavior in our data. With one notable exception, heavy-tailed behavior was only detected in the dry seasons of two of the nine rivers. Furthermore, the magnitudes associated with these power law tails were well within the range of routine wet season flow values. Hence, we presented a cautionary note on the unscrupulous application of heavy-tailed models in the absence of adequate evidence from data. Furthermore, we found that some rivers' seasonal flows follow distributions in the same family, while other rivers' seasonal flows follow distributions from different families. Some possible physical explanations were discussed to explain this situation. An enhanced distributional model, involving a stochastic process with time-varying parameters, was suggested as a possible improvement to the methods presented here.

[47] Finally, the apparent discrepancy in the parameter estimates of α between *Anderson and Meershaert* [1998] and the present paper is not trivial. Noting the different temporal resolutions used (monthly versus daily) in the analyses, it is reasonable to expect that the averaging process used to obtain the lower-resolution data has subdued some of the extremes observed in the higher-resolution data. This in turn would increase the estimate of α in the lower-resolution data. A simulation using IID random variables proved this. So far, however, there is no general theory to compute α for averaged, coarse-resolution, data based on the power law exponent of its high-resolution counterpart and the parameter for smoothing. The development of such a theory would be an important research task in the future.

[48] **Acknowledgments.** We thank two anonymous reviewers and the Associate Editor for many constructive comments. We thank A. Clauset and colleagues for providing code for the estimation of MLEs for power law parameters b and α. The code was downloaded from http://tuvalu.santafe.edu/~aaronc/powerlaws/. This work is supported in part by NSF grants CMMI-1031958 and 0826119.

References

Aban, I. B., M. M. Meerschaert, and A. K. Panorska (2006), Parameter estimation of the truncated Pareto distribution, *J. Am. Stat. Assoc.*, *101*(473), 270–277.

Anderson, P. L., and M. M. Meershaert (1998), Modeling river flows with heavy tails, *Water Resour. Res.*, *34*(9), 2271–2280.

Atlas, D., D. Rosenfeld, and D. A. Short (1990), The estimation of convective rainfall by area integrals: 1. The theoretical and empirical basis, *J. Geophys. Res.*, *95*(D3), 2153–2160.

Benyahya, L., D. Caissie, F. Ashkar, N. El-Jabi, and M. Satish (2009), Comparison of the annual minimum flow and the deficit below threshold approaches: Case study for the province of New Brunswick, Canada, *Can. J. Civ. Eng.*, *36*(9), 1421–1434.

Bernardara, P., D. Schertzer, E. Sauquest, I. Tchiguirinskaia, and M. Lang (2008), The flood probability distribution tail: How heavy is it?, *Stochastic Environ. Res. Risk Assess.*, *22*, 107–122.

Black, A. R., J. S. Rowan, R. W. Duck, O. M. Bragg, and B. E. Clelland (2005), DHRAM: A method for classifying river flow regime alterations for the EC Water Framework Directive, *Aquat. Conserv. Mar. Freshwater Ecosyst.*, *15*, 427–446.

Carreau, J., P. Naveau, and E. Sauquet (2009), A statistical rainfall-runoff mixture model with heavy-tailed components, *Water Resour. Res.*, *45*, W10437, doi:10.1029/2009WR007880.

Castellarin, A., R. M. Vogel, and A. Brath (2004), A stochastic index flow model of flow duration curves, *Water Resour. Res.*, *40*, W03104, doi:10.1029/2003WR002524.

Clauset, A., C. R. Shalizi, and M. E. J. Newman (2009), Power-law distributions in empirical data, *SIAM Rev.*, *51*(4), 661–703, doi:10.1137/070710111.

De Domenico, M., and V. Latora (2011), Scaling and universality in river flow dynamics, *Europhys. Lett.*, *94*, 58002, doi:10.1209/0295-5075/94/58002.

DeGroot, M. H., and M. J. Schervish (2002), *Probability and Statistics*, 3rd ed., Addison-Wesley, Boston, Mass.

Elshorbagy, A., S. P. Simonovic, and U. S. Panu (2002), Noise reduction in chaotic hydrologic time series: Facts and doubts, *J. Hydrol.*, *256*, 147–165.

Fernandes, W., M. Naghettini, and R. Loschi (2010), A Bayesian approach for estimating extreme flood probabilities with upper-bounded distribution functions, *Stochastic Environ. Res. Risk Assess.*, *24*(8), 1127–1143.

Fernández, B., and J. D. Salas (1999), Return period and risk of hydrologic events. I: Mathematical formulation, *J. Hydrol. Eng.*, *4*(4), 297–307.

Foster, J., and M. Bevis (2003), Lognormal distribution of precipitable water in Hawaii, *Geochem. Geophys. Geosyst.*, *4*(7), 1065, doi:10.1029/2002GC000478.

Gao, J. B., and I. Rubin (2001a), Multiplicative multifractal modeling of long-range-dependent network traffic, *Int. J. Commun. Systems*, *14*, 783–801.

Gao, J. B., and I. Rubin (2001b), Multifractal modeling of counting processes of long-range-dependent network traffic, *Comput. Commun.*, *24*, 1400–1410.

Gao, J. B., Y. H. Cao, W. W. Tung, and J. Hu (2007), *Multiscale Analysis of Complex Time Series: Integration of Chaos and Random Fractal Theory, and Beyond*, John Wiley, Hoboken, N. J.

Gao, J. B., H. Sultan, J. Hu, and W. Tung (2010), Denoising nonlinear time series by adaptive filtering and wavelet shrinkage: A comparison, *IEEE Signal Process. Lett.*, *17*, 237–240.

Gao, J. B., J. Hu, and W. W. Tung (2011), Facilitating joint chaos and fractal analysis of biosignals through nonlinear adaptive filtering, *PLoS ONE*, *6*(9), e24331, doi:10.1371/journal.pone.0024331.

Goldstein, M. L., S. A. Morris, and G. G. Yen (2004), Problems with fitting to the power-law distribution, *Eur. Phys. J. B.*, *41*(2), 255–258.

Hu, J., J. B. Gao, and X. S. Wang (2009a), Multifractal analysis of sunspot time series: The effects of the 11-year cycle and Fourier truncation, *J. Stat. Mech., 2009*, P02066, doi:10.1088/1742-5468/2009/02/P02066.

Hu, J., W. W. Tung, and J. B. Gao (2009b), A new way to model nonstationary sea clutter, *IEEE Signal Process. Lett.*, *16*(2), 129–132.

Iacobellis, V. (2008), Probabilistic model for the estimation of *T* year flow duration curves, *Water Resour. Res.*, *44*, W02413, doi:10.1029/2006WR005400.

Katz, R. W., M. B. Parlange, and P. Naveau (2002), Statistics of extremes in hydrology, *Adv. Water Resour.*, *25*, 1287–1304.

Kroll, C. N., and M. Vogel (2002), Probability distribution of low streamflow series in the United States, *J. Hydrol. Eng.*, *7*, 137–146.

Laio, F. (2004), Cramer-von Mises and Anderson-Darling goodness of fit tests for extreme value distributions with unknown parameters, *Water Resour. Res.*, *40*, W09308, doi:10.1029/2004WR003204.

Lampariello, F. (2000), On the use of the Kolmogorov-Smirnov statistical test for immunofluorescence histogram comparison, *Cytometry*, *39*, 179–188.

Li, L. J., L. Zhang, H. Wang, J. Wang, J. W. Yang, D. J. Jiang, J. Y. Li, and D. Y. Qin (2007), Assessing the impact of climate variability and human activities on streamflow from the Wuding River basin in China, *Hydrol. Processes*, *21*, 3485–3491.

Lopez, R. E. (1977), The lognormal distribution and cumulus cloud populations, *Mon. Weather Rev.*, *105*, 865–872.

Mandelbrot, B. (1963), The variation of certain speculative prices, *J. Bus.*, *36*, 394–419.

Montroll, E., and M. Shlesinger (1982), On 1/*f* noise and other distributions with long tails (log-normal distribution, Levy distribution, Pareto distribution, scale-invariant process), *Proc. Natl. Acad. Sci. U. S. A.*, *79*, 3380–3383.

Montroll, E., and M. Shlesinger (1983), Maximum entropy formalism, fractals, scaling phenomena, and 1/*f* noise: A tale of tails, *J. Stat. Phys.*, *32*, 209–230.

Morrison, J. E., and J. A. Smith (2002), Stochastic modeling of flood peaks using the generalized extreme value distribution, *Water Resour. Res.*, *38*(12), 1305, doi:10.1029/2001WR000502.

Newman, M. (2005), Power laws, Pareto distributions and Zipf's law, *Contemp. Phys.*, *46*(5), 323–351.

Ochou, A. D., A. Nzeukou, and H. Sauvageot (2007), Parametrization of drop size distribution with rain rate, *Atmos. Res.*, *84*, 58–66.

Pandey, G., S. Lovejoy, and D. Schertzer (1998), Multifractal analysis of daily river flows including extremes for basins of five to two million square kilometers, one day to 75 years, *J. Hydrol.*, *208*, 62–81.

Poff, N. L., J. D. Allan, M. B. Bain, J. R. Karr, K. L. Prestegaard, B. D. Richter, R. E. Sparks, and J. C. Stromberg (1997), The natural flow regime, *BioScience*, *47*(11), 769–784.

Sharwar, M. M., K. Sangil, and P. Jeong-Soo (2011), Beta-kappa distribution and its application to hydrologic events, *Stochastic Environ. Res. Risk Assess.*, *25*(7), 897–911.

Sivakumar, B. (2004), Chaos theory in geophysics: Past, present and future, *Chaos Solitons Fractals*, *19*, 441–462.

Smakhtin, V. Y. (2001), Low flow hydrology: A review, *J. Hydrol.*, *240*, 147–186.

Smakhtin, V. Y., D. A. Hughes, and E. Creuse-Naudin (1997), Regionalization of daily flow characteristics in part of the Eastern Cape, South Africa, *Hydrol. Sci.*, *42*(6), 919–936.

Stedinger, J. R. (1980), Fitting log normal distributions to hydrologic data, *Water Resour. Res.*, *16*(3), 481–490.

Tesfaye, Y. G., M. M. Meerschaert, and P. L. Anderson (2006), Identification of periodic autoregressive moving average models and their application to the modeling of river flows, *Water Resour. Res.*, *42*, W01419, doi:10.1029/2004WR003772.

Tung, W. W., M. W. Moncrieff, and J. B. Gao (2004), A systemic view of the multiscale tropical deep convective variability over the tropical western-Pacific warm pool, *J. Clim.*, *17*, 2736–2751.

Tung, W. W., J. B. Gao, J. Hu, and L. Yang (2011), Detecting chaos in heavy-noise environments, *Phys. Rev. E*, *83*, 046210.

Um, M. J., W. Cho, and J. H. Heo (2010), A comparative study of the adaptive choice of thresholds in extreme hydrologic events, *Stochastic Environ. Res. Risk Assess.*, *24*(5), 611–623.

Villarini, G., J. A. Smith, F. Serinaldi, and A. A. Ntelekos (2011), Analyses of seasonal and annual maximum daily discharge records for central Europe, *J. Hydrol.*, *399*, 299–312.

Vogel, R. M., and N. M. Fennessey (1994), Flow-duration curves. I. new interpretations and confidence intervals, *J. Water Resour. Plann. Manage.*, *120*(4), 485–504.

Vörösmarty, C. J., and D. Sahagian (2000), Anthropogenic disturbance of the terrestrial water cycle, *BioScience*, *50*(9), 753–765.

Wang, W., J. K. Vrijling, P. H. A. J. M. Van Gelder, and J. Ma (2006), Testing for nonlinearity of streamflow processes at different timescales, *J. Hydrol.*, *322*, 247–268.

Zhang, Q., C. Y. Xu, Y. Q. Chen, and Z. G. Yu (2008), Multifractal detrended fluctuation analysis of streamflow series of the Yangtze River basin, China, *Hydrol. Processes*, *22*, 4997–5003.

Zhou, Y. H., J. B. Gao, K. D. White, I. Merk, and K. Yao (2004), Perceptual dominance time distributions in multistable visual perception, *Biol. Cybern.*, *90*, 256–263.