Spatial scaling in a changing climate: A hierarchical bayesian model for non-stationary multi-site annual maximum and monthly streamflow

Carlos H.R. Lima *, 1, Upmanu Lall

Columbia University, Water Center, 500 W, 120th Street, Office 842, New York, NY 10027, USA

ARTICLE INFO

Article history:
Received 5 April 2009
Received in revised form 2 November 2009
Accepted 29 December 2009
Available online xxxxx

This manuscript was handled by A. Bardossy, Editor-in-Chief, with the assistance of Ercan Kahya, Associate Editor

Keywords:
Hierarchical Bayesian models
Drainage area scaling
Regionalization
Streamflow spatial scaling
Flood frequency analysis
Flood flow regionalization

SUMMARY

Several studies have shown that statistics of streamflow time series, in particular empirical moments, scale with physical properties of the drainage basin, such as the catchment area. Those scaling laws have been extensively used to estimate statistics of streamflow series at ungauged sites. The role of climate variability and change has not been considered in such models. Further, most studies are based on classical statistics, where parameter uncertainties are usually neglected or not formally considered. In this paper we develop and apply hierarchical Bayesian models, to both assess regional and at-site trends in time in a spatial scaling framework, and simultaneously provide a rigorous framework for assessing and reducing parameter and model uncertainties. The models are tested with reconstructed natural inflow series from over 40 hydropower sites in Brazil with catchments areas varying from 2588 to 823,555 km². Both annual maximum flood series and monthly streamflow are considered. Cross-validated results show that the Hierarchical Bayesian models are able to skillfully estimate monthly and flood flow probability distribution parameters for sites that were not used in model fitting. The models developed can be used to provide record augmentation at sites that have short records, or to estimate flow at ungauged sites, even in the absence of an assumption of time stationarity. Since model uncertainties are accounted for, the precision of the estimates can be quantified and hypotheses tests for regional and at-site trends can be formally made. A formal inclusion of climate predictors to facilitate seasonal forecasting or climate change scenario development is also feasible. This is indicated, but not developed here.

© 2009 Elsevier B.V. All rights reserved.

Introduction

Regionalization or regional analysis of hydro-climatological variables, such as streamflow, rainfall, evaporation and their associated statistics (e.g. means, standard deviation, flood quantiles, etc.), has been an active area of research over the last 40 years. The understanding of the spatial variability of these statistics is important for hydro-climatological time series and in the water resources management. For instance, flood management and design of flood control structures (e.g. dams, bridges, spillways, culverts) usually require the estimate of low exceedance probabilities, e.g. 1% for the 100-year flood quantile, which in turn demands a sufficient amount of data (no less than 100 years of streamflow record for this case) for reliable estimates. Since the desired amount of data is rarely available, one wants to use hydroclimatic information of similar and nearby sites to produce a better, more reliable estimate of the quantiles associated with those low probabilities of occurrence (Stedinger et al., 1993). The recognition that climate is inherently variable and changing also brings the question of how such changes should be modeled as part of a formal non-stationary analysis (Milly et al., 2008; Jain and Lall, 2001). The work presented here addresses aspects of spatial scaling, nonstationarity and uncertainty analysis from a regional perspective working with multiple time series of monthly flows and annual maximum flow.

At sites with no record of streamflow data (ungauged sites), regional analysis is used to estimate the variable of interest (e.g. the 100-year flood quantile) at sites with historical data of streamflow available and then relate the estimate with physiographic and geomorphologic features of the associated region and catchment basin. Common features used include drainage area, channel slope and length, vegetation cover, soil properties, altitude, as well as historical and paleo information (e.g. Thomas and Benson, 1970; Stedinger and Cohn, 1986; Martins and Stedinger, 2001). A statistical model correlating explanatory and response variables is then obtained in order to estimate the desired statistics at the ungauged site (the problem of prediction in ungauged basins – PUB, see Gupta et al., 2007), where the only information available is related to the explanatory variables. Regionalization can also be used to improve parameter estimates (e.g. the index flood method) and...
time series augmentation, where only short records of the desired variable are present (Salas et al., 1980; Stedinger et al., 1993).

Several attempts have been made to identify spatial scaling attributes of streamflow statistics and catchment physical properties (e.g. Thomas and Benson, 1970; Riggs, 1973; Pandey et al., 1998; Vogel and Sankarasubramanian, 2000; Yue and Gan, 2004; Koscielny-Bunde et al., 2006) or simply to improve the methods used to link response variables and predictors in current statistical models (e.g. Stedinger and Tasker, 1985; Tasker and Stedinger, 1989; Kroll and Stedinger, 1998; Pandey and Nguyen, 1999). In particular, annual mean flow and annual peak floods for given return periods have been known for a long time to scale as power laws with catchment area (e.g. Benson, 1962; Thomas and Benson, 1970; Alexander, 1972), which is the most common variable used in regionalization due to its availability (one can easily obtain the drainage area for almost any streamflow site) and reliability (the estimates are very precise). The original scaling relationships have been also linked to the literature on multi-fractals where the scaling exponents vary by the moment order (Gupta and Waymire, 1990; Smith, 1992; Gupta et al., 1994; Becker and Braun, 1999; Vogel and Sankarasubramanian, 2000; Sivapalan et al., 2002; Yue and Gan, 2004). Poveda et al. (2007) used the long-term water balance equation to estimate the mean annual flow and then made use of the power law to estimate the mean and standard deviation of annual floods which in turn were used to calculate annual floods for any given return period. A more comprehensive review of the so-called scaling theory within a non-linear geophysical framework is presented in Gupta et al. (2007), which also discuss the use of power laws under global climate changes. However, in none of the key papers in this area has a method for assessing temporal variations in parameters that may be related to climate or other factors been discussed. We address this issue.

Following the principle of introduction of more 'structure' into models proposed by the National Research Council (National Research Council, 1988), we develop here a hierarchical Bayesian model (Gelman et al., 2004), where parameter uncertainties are fully accounted into model outputs and information of different sources is used in order to shrink those uncertainties and improve the model reliability. With few exceptions (e.g. Júnior et al., 2005; Poveda et al., 2007; Kwon et al., 2008), regionalization has been based on classical statistics, where parameters are assumed stationary in time and space and their associated uncertainties are usually neglected or rely on asymptotic normality assumptions and are not fully accounted in model outputs. With the framework proposed here, we are able to include more information and better estimate time varying parameters of frequency distributions of annual maximum as well as flood quantiles and monthly streamflows at gauged and ungauged sites, making a significant contribution to the PUB problem. This paper is organized as follows: in the next Section we describe the monthly and annual streamflow data. In Section “Spatial scaling and a bayesian model for the parameters of the probability distribution of annual maximum flood series” we describe a Bayesian model for the parameters of a probability distribution fit to annual maximum flood series. A hierarchical Bayesian model for time varying parameters of the annual maximum series is presented in Section “Hierarchical bayesian modeling of non-stationarity in the scaling of annual maximum flood series”. Finally in Section “Hierarchical bayesian modeling of non-stationary monthly streamflow series” we consider a hierarchical Bayesian model for monthly streamflow series.

River discharge data

Daily streamflow data of 44 hydropower sites in Brazil (location shown in Fig. 1) are provided by the System National Operator (ONS), which is the Brazilian institution responsible for defining operation rules and strategies to maximize the electrical energy production across thermal and hydro plants. As displayed in Fig. 1, 32 sites are located in nested basins, which together form one of the largest basins (in terms of water flow) in South America, the Paraná basin. The streamflow time series cover the January 1931–December 2001 period and span a large range of power capacities (from 80 MW to 14000 MW) and catchment areas (from 322 to 823,555 km²). Most studies of scaling in annual flood frequencies have been limited to catchment areas of the order of 5000 km² (e.g. Gupta et al., 1994). To our knowledge, Poveda et al. (2007) was the first article to investigate annual flood scaling statistics for catchment areas of the order of 1 million km². Most such analyses consider a nested basin structure, under the assumption of a homogeneous climate/rainfall distribution over the region, and topographic/geomorphic controls as the important drivers of the scaling relationship of interest between flow and drainage area. The framework we present here, and the data set used are not limited to this structure. We can allow both spatial and temporal variation in parameters. The latter is formally explored, while the former can be diagnosed through an analysis of the site by site variation in model parameters.

The daily series of streamflow used here are reconstructed natural time series, i.e., estimated river flow after accounting for estimated water use (e.g. reservoir operation, water withdrawals) upstream of the gauge. The annual maximum series are calculated by taking the maximum daily flow observed for each year of the record. Not all sites have a full record of available data. Fig. 2 shows the percentage of hydropower reservoirs with available daily flow data as a function of year. Note that only beyond 1973 one has a complete set of sites with daily streamflow data. Extension of these records to fill in missing values with estimates of the associated uncertainty is one of the motivations of the proposed models.

Monthly inflow series of 45 hydropower sites in Brazil are also provided by ONS. These series cover the 1931–2006 period and are also reconstructed time series. They have been verified and revised by the Brazilian National Water Agency (ANA) and do not have any
missing values. The geographical location of these sites is displayed in Fig. 1. The associated drainage area varies from 2588 to 823,555 km².

Most of the catchments displayed in Fig. 1 have similar patterns of rainfall seasonality and climate. The rainy season is driven mainly by the South Atlantic Convergence zone with remote forcings from the Tropical Pacific and South Atlantic sea surface temperatures (SST). A detailed description of the climate and teleconnection patterns associated with the rainfall regime across Brazil can be found in Ropelewski and Halpert (1987), Nogues-Paegle and Mo (1997), Grimm (2004), Vera et al. (2006) and references therein. Geomorphological attributes of the correspondent drainage catchments are very diverse, and we do not attempt to investigate them here. We only consider the drainage area, which spans three orders of magnitude. It is well known that the annual mean flow scales with drainage area. Menabde and Sivapalan (2001) show through an idealized physical model that flood events also scale with the catchment area. Gupta et al. (2007) review some related work and argue that random self-similarity in drainage networks produces power laws in floods on event time scales.

Spatial scaling and a Bayesian model for the parameters of the probability distribution of annual maximum flood series

Simple spatial scaling for an arbitrary random field $Y(x)$ is defined Gupta and Waymire (1990) as:

$$E[Y_h^k] = \lambda \cdot E[Y_1^h]$$

where $\lambda > 0$ is a spatial scale parameter.

This can be expressed as:

$$\log E[Y_h^k] = h \lambda \log \lambda + \log E[Y_1^h]$$

where $\lambda$ is the scaling exponent, $h$ the moment order and $Y_1$ the re-scaled random field:

$$Y_h^k(x) = Y(hx).$$

Gupta and Waymire (1990) showed that the log–log linearity expressed in (2) with respect to the drainage area holds for several instantaneous streamflow data series across the United States, but the linear slope property $h \rightarrow h \lambda$ does not. In fact, the slope of (2) was found to be a nonlinear, concave function of the moment order $h$, suggesting a multiscaling process. Gupta and Dawdy (1995) also observed both simple and multiscaling in regional annual flood frequencies. Several authors (e.g. Smith, 1992; Gupta et al., 1994; Becker and Braun, 1999; Vogel and Sankarasubramanian, 2000; Sivapalan et al., 2002; Yue and Gan, 2004) have investigated such scaling behaviors for other hydrological data. We refer to Gupta et al. (2007) for a comprehensive review of related works.
The Gumbel distribution and the scaling law of its parameters

We use the Gumbel distribution, which is often employed in frequency analysis of annual maximum flood series (Stedinger et al., 1993) to exemplify the Bayesian model presented. Let

\[ q_i \sim \text{Gumbel}(a, b), \]

where its distribution function is given by:

\[ F(q_i | a, b) = e^{-e^{\frac{-q_i}{b}}}, \]

and \( q_i \) is the at-site annual maximum at year \( i \) and \( a \) and \( b \) are, respectively, the location and scale parameters. These parameters are related to the moments of the distribution as (Stedinger et al., 1993):

\[ \mu = \bar{q} - \gamma \frac{\sigma}{2}, \]

\[ \sigma = \frac{\sqrt{6}\sigma}{\pi}, \]

where \( \bar{q} \) and \( \sigma^2 \) are the sample mean and variance of \( q \) across years and \( \gamma \approx 0.5772 \) is the Euler’s constant.

Panels of Fig. 3a and b display the log-log linear relationship of the first two moments (mean and standard deviation) of the annual maximum series of 35 hydropower sites and their corresponding drainage area. The linear relationship in (6) and (7) suggest that the location and scale parameters of the Gumbel distribution will also scale with drainage area, as shown in Panels of Fig. 3c and d. Using a physical based model, Menabde and Sivapalan (2001) also show similar scaling laws for the Gumbel parameters.

Note that the Gumbel distribution is used here as an illustrative example. The actual form of the annual flood frequency distribution remains a challenge in hydrology and its demonstration from physical processes is still a fundamental unsolved problem. The Bayesian model presented in the next section for the Gumbel distribution can also be applied to estimation with other distributions, such as the Log Normal, Weibull, GEV and Generalized Pareto.

A bayesian model for the Gumbel parameters

We first assume that the annual maximum series follow a conditional independent Gumbel distributions given the location \( a_k \) and scale \( b_k \) parameters of site \( k \):

\[ q_{ik} \sim \text{Gumbel}(a_k, b_k), \]

One way to estimate the Gumbel (or other) distribution parameters for ungauged sites, i.e., sites with little or no information on annual maximum floods but with drainage area information, is to
obtain the first two product moments of its distribution through some equation that relates these distribution parameters to the drainage area for sites with data (for instance, obtaining from Panels of Fig. 3a and b). These “regression” based estimates can then be used in (6) and (7) to obtain estimates for the location and scale parameters at the unaged sites. Usually, the uncertainty in the estimates of the at-site moments for the sites with records (potentially of unequal lengths) or in the regression vs drainage area is not formally considered or transmitted to the subsequent estimation process. Also, the changes of these parameters with time due to climate or land use are not formally integrated.

Under a Bayesian framework, we assume that the parameters in (8) follow a probability distribution, i.e., their prior distribution.

Since empirical (Panels of Fig. 3c and d) as well as physical model based (Menabde and Sivapalan, 2001) evidence indicates that both Gumbel parameters scale with the drainage area, we assume that the prior distribution of location and scale parameters also follows a log–log linear relationship with respect to the drainage area:

\begin{align}
p(\log(a_k)) & \sim N(\alpha_0 + \alpha x(k), \sigma_\alpha^2) \tag{9} \\
p(\log(b_k)) & \sim N(\beta_0 + \beta x(k), \sigma_\beta^2) \tag{10}
\end{align}

where \(k\) refers to streamflow gauge \(k\) and \(x(k)\) is the zero mean logarithmic area, defined as \(x(k) = \log(A(k)) - \log(A)\), where \(A(k)\) is the catchment area of site \(k\) and \(\log(A)\) represents the average of the logarithmic of the catchment areas across all sites. The reason for a zero mean predictor is a reparameterization procedure in order to reduce the correlation between the regression parameters in (9) and (10) and facilitate the convergence of their estimates (Gilks and Roberts, 1995).

Usually there is enough information (35 data points correspondent to 35 streamflow gauges for the model tested here) to estimate the six parameters in (9) and (10). Consequently, non-informative prior distributions (independent, uniform) are adopted for the parameters in (9) and (10) as suggested in the literature (Gelman et al., 2004; Gelman, 2005):

\[ p(\alpha_0, \beta_0, \beta_1, \sigma_\alpha, \sigma_\beta) \propto 1 \]  

Bayes’ rule allows one to develop the posterior density for the model parameters as:

\[ p(A|q) = \frac{p(A, q)}{p(q)} = p(q|A) \cdot p(A) \]  

where \(A = [a_k, b_k, \alpha_0, \beta_0, \beta_1, \sigma_\alpha, \sigma_\beta]\), \(k = 1, \ldots, K\) refers to the entire set of parameters, \(K\) is the total number of streamflow sites, \(p(A)\) is referred as the prior distribution of the parameters and \(p(q|A)\) is the likelihood function of the data given by:

\[ p(q|A) = \prod_{k=1}^{K} \prod_{j=1}^{n_k} \text{Gumbel}(q_{ik} | a_k, b_k) \]  

where \(n_k\) is the number of years with available data for site \(k\). Note that this allows one to explicitly consider the unequal sample size available across the sites and to formally consider it in the estimation process.

![Fig. 5. Log–log relationship of annual maximum series and drainage area for some selected years of the record. The variable along the x-axis is defined as in Fig. 3. Units are m\(^3\)/s for annual maximum series and km\(^2\) for drainage area.](image)
Substituting the prior of the parameters as defined in (9)–(11), and the likelihood function (13) into (12), yields the joint posterior distribution of the parameters:

\[
p(A|q) \propto \prod_{k=1}^{K} \prod_{i=1}^{n_k} \text{Gumbel}(q_{ik}|a_k, b_k) \cdot N(\log(a_k)|x_0) \\
+ \sigma_i \cdot N(\log(b_k)|\beta_0 + \beta_1 \cdot x(k), \sigma^2)
\]  \( (14) \)

Eq. (14) involves the estimation of several parameters, with non-conjugate prior distributions for the Gumbel scale and location parameters and for the regression parameters. The integral over all parameters can not be directly solved. In this case, we have to turn on to other methods. We adopt here the widely used Markov Chain Monte Carlo (MCMC) method to draw values of the set of parameters from their posterior distribution (14). In particular, we combine the Gibbs sampler and the Metropolis algorithm (Gelman et al., 2004) for simulating from (14). We apply Gibbs sampler for the regression parameters, since a closed form of the conditional posterior distribution (normal distribution in this case) can be easily obtained given the uniform prior distribution (Gelman et al., 2004) adopted. The Metropolis algorithm is used to obtain samples from the conditional posterior distribution of the location and scale parameters, since there is no closed form for this distribution. A MCMC simulation as described above is run with five chains to verify the convergence of the results (or the mixing) based on the methodology suggested by Gelman et al. (2004).

In order to verify the skill of the model in reproducing the data (predictive check), 1000 simulations of the joint posterior distribution of the regression parameters in (9) and (10) were drawn and applied to nine randomly selected streamflow sites that were not used in the MCMC simulation. Fig. 4 shows the posterior distribution of the 100-year flood (i.e., the 1 – 1/100 = 0.99 quantile in Eq. (5)) along with the estimates obtaining after fitting a Gumbel distribution to the observed annual maximum series. A general agreement is obtained for all sites. The estimated 100-year flood based on site data (vertical bar in Fig. 4) lies well within the distribution of the 100-year flood based on the MCMC estimates.

In the context of PUB, the model proposed here is able to predict the (posterior) distribution rather than point estimates of any annual maximum statistic given only the drainage area of the ungauged sites. Doing that, we are able to provide the uncertainty band of our estimates (the 100-year flood in case of Fig. 4) after simultaneously accounting for (i) the uncertainty in the Gumbel distribution parameters of the sites with available data and (ii) uncertainty from the scaling law regression.

Hierarchical bayesian modeling considering nonstationarity in the scaling of annual maximum flood series

A key problem that has been highlighted recently is that anthropogenic climate change and land use, as well as natural climate variability at inter-annual and decadal time scales lead to nonstationarity in the probability distribution of floods and other hydrologic variables. If these relations are temporarily variable, then reconstructing past historical data at ungauged locations becomes much more challenging, since the statistical relationship that should be used for the purpose may itself have parameters that are changing with time. In this Section we consider the potential importance of this issue for the Brazilian data, and propose and apply a methodology that allows for a formal consideration of these time varying relationships, i.e., a non-stationary model. Fig. 5 shows the log–log scaling law of annual maximum series and catchment area for selected years of the record. This motivates the modeling of annual maximum series through time varying scaling coefficients, while minimizing the increase in uncertainty of estimation in the process:

\[
\log(q_{ik}) \sim N(\phi_0 + \phi_1 x(k), \tau^2_i)
\]  \( (15) \)

where \( k \) and \( i \) represent site and year, respectively.

---

The limited amount of data to estimate the regression parameters in (15) leads to low degrees of freedom and high uncertainty in the estimates. For instance, Fig. 2 shows that in the beginning of the record (between 1930 and 1940) one had less than 60% (or about 25 reservoirs) of the total number of hydropower sites with available data. In order to pool more information for estimating the time varying parameters in (15) and consequently reduce their uncertainty, one can assume that those parameters are drawn from a common distribution, which is represented by their long time average (the distribution that would apply under stationarity), but with some additional variation that may or may not be systematic with time. The amount of additional variation to allow is a parameter of a hierarchical model. In effect, one can think of a model that assumes the same relationship holds across all years as a fully pooled regression – all years contribute in the same way to the regression and are assumed to have the same underlying parameters. Conversely, if scaling parameters could be estimated separately for each year, then one would assume that each year’s scaling represents a separate process. The Hierarchical Bayesian model solves for the amount of variation to allow from the fully pooled model so that an appropriate degree of pooling across years is allowed and departures from the underlying pooled model can be appropriately recognized with a quantification of the associated uncertainty. These departures may be purely random in time due perhaps to outliers in data, or they may exhibit systematic variation in time, in which case a specific hypothesis for the form of the nonstationarity can be formulated and explored. Of course, in the process one would have an equation for each year to apply for filling in values at ungauged locations that recognizes the changing climate conditions that may apply to that year. Mathematically, we can proceed by defining the following prior distribution:

$$\begin{pmatrix} \phi_{0i} \\ \phi_{1i} \end{pmatrix} \sim N \left( \begin{pmatrix} \bar{\phi}_0 \\ \bar{\phi}_1 \end{pmatrix}, \Sigma \right)$$

(16)

where $\bar{\phi}_0$ and $\bar{\phi}_1$ can be considered to be the parameters that apply across years, and $\Sigma$ is a covariance matrix of these parameters.

For simplicity, since there is no a priori information about the variance term in (15), we just assume a uniform prior distribution:

$$p(\tau_i) \propto 1.$$  

(17)

Usually there are enough data available to estimate the mean and covariance matrix in (16), so a common choice of prior distributions can be independent uniform priors (in the case of $\Sigma$, a Jeffreys prior, see Gelman et al., 2004) for all parameters. However, initial analysis with our data showed that the off diagonal elements of $\Sigma$ and the
variance of $\phi_1$, are relatively close to zero (order of $10^{-4}$), which makes MCMC convergence difficult (Gilks and Roberts, 1995). Since finding uniform (or non-informative) priors for such type of covariance matrices is still a topic of research in Bayesian statistics (see for instance, Gilks and Roberts (1995) or the suggestion in Gelman (2005) to generalize his recently proposed prior distribution for variance parameters to covariance matrices), we adopt here conjugate prior distributions with hyperparameters $(A_0, v_0, \phi_0, \phi_0, k_0)$ estimated from data:

$$\Sigma \sim \text{Inv-Wishart}_{v_0}(A_0)$$

$$\left(\begin{array}{c}
\phi_0 \\
\phi_1
\end{array}\right) \sim N\left(\begin{array}{c}
\phi_0 \\
\phi_1
\end{array}; \Sigma/k_0\right)$$

where $v_0$ and $A_0$ are, respectively, the degrees of freedom and the inverse scale matrix of the inverse Wishart distribution.

Combining prior distributions and the likelihood function (15) yields the joint posterior distribution of the complete set of parameters $\Phi$:

$$p(\Phi|q) \propto \prod_{k=1}^{n} N(\log(q_k)|\phi_0 + \phi_1x(k), \Sigma) \cdot N\left(\begin{array}{c}
\phi_0 \\
\phi_1
\end{array}; \Sigma/k_0\right)$$

Posterior samples of parameters from (20) are drawn using the Gibbs sampler algorithm (Gelman et al., 2004). The posterior parameters $\phi_0$ and $\phi_1$ are a weighted average of the likelihood function and the prior distribution, resulting in a multivariate normal distribution (Gelman et al., 2004, pp. 86). The posterior distributions of $\phi_0$, $\phi_1$, and $\Sigma$ are also conjugate distributions (i.e., they are in the same family of the prior distribution). Gelman et al. (2004, pp. 87–88) shows how to obtain those conditional posteriors as weighted averages of the prior distribution and likelihood function.

Fig. 6 displays the Bayesian estimates (expected value of the posterior distribution) of the time varying intercept and slope parameters along with the 95% interval. Pooled estimates (i.e., maximum likelihood estimates using the complete data and assuming

\[ y = 6.1 + 0.79 \times x \]

\[ y = 6.8 + 0.81 \times x \]

\[ y = 7.4 + 0.88 \times x \]

\[ y = 6.9 + 0.86 \times x \]

\[ y = 7.4 + 0.86 \times x \]

\[ y = 6.8 + 0.86 \times x \]

\[ y = 7.4 + 0.88 \times x \]

\[ y = 6.8 + 0.86 \times x \]

\[ y = 7.4 + 0.88 \times x \]

\[ y = 6.8 + 0.86 \times x \]
Hierarchical bayesian modeling of non-stationary monthly streamflow series

Estimates of past monthly flows at ungauged sites or augmentation of time series at short record sites are of very interest for several reasons. For instance, in calibrating optimization models of hydro energy production one often needs monthly series of inflow into the hydropower reservoirs within the system. The design of water reservoirs also requires estimates of past monthly flows in order to define operational policies for water releases and storage, in particular for multi-year regularization reservoirs.

Empirical data (Fig. 8) from 36 streamflow sites in Brazil (spatial location displayed in Fig. 1) show strong evidences that monthly flows also scales with drainage area for sites with similar seasonal patterns. The scaling exponents for monthly flows with drainage area are consistently less than one. A reviewer noted that if we consider that the monthly flow volume $V$ is given as the product of the average runoff $R$ per unit area and the drainage area, $A$, then a slope of 1 is expected for the log $V$ vs log $A$ relation under the assumption that $R$ is independent of $A$. Since this exponent is less than unity (estimated slopes typically fall in the range $\{0.8, 0.9\}$), an investigation into why $R$ apparently decreases with $A$ is necessary. It is likely that heterogeneity in rainfall or the coverage of area by storms decreases as the area increases, and hence the average runoff produced per storm and by accumulation across storms, for monthly flow decreases a bit as the underlying area increases.

The validity of this conjecture needs to be assessed. Inability to access highly detailed time series of event or even monthly rainfall to establish spatial coverage prevented us from exploring this direction. A mechanistic model could indeed be used with synthetic rainfall coverage and other variations to explore this idea.

Similarly to the model for the annual maximum as defined in (15), we can define a stochastic model for the monthly flow $y$:

$$\log(y_{jk}) \sim N(\theta_0 + \theta_1 x(k), \sigma^2_0)$$

(21)

where $y_{jk}$ is the streamflow of site $k$ at month $j$ of year $i$.

Fig. 9. Time average (lines with dots) intercept and slope of scaling law parameters for the monthly flow (Eqs. (21) and (22)). The black lines show the fitting of one-harmonic Fourier function as defined in Eqs. (25) and (26).

In the first hierarchy of the model, we shrink the monthly intercept and slope parameters in (21) towards a common distribution across years as was done in the previous Section for annual maximum flows:

$$
\begin{pmatrix}
\bar{\theta}_0 \\
\bar{\theta}_1
\end{pmatrix}
\sim N
\begin{pmatrix}
\bar{\theta}_0 \\
\bar{\theta}_1
\end{pmatrix}, \Sigma
$$

(22)

For simplicity, we use a common approach in the literature (Gelman et al., 2004) and define uniform prior distribution for the scale parameters:

$$
p(\sigma_i) \propto 1.
$$

(23)

Empirical analysis (Fig. 9) also shows that the time average (across years) of the intercept and slope parameters in (21) follows a well defined seasonal cycle. In order to maintain this seasonal behavior, one more step of shrinkage is possible:

$$
\begin{pmatrix}
\bar{\theta}_0 \\
\bar{\theta}_1
\end{pmatrix}
\sim N
\begin{pmatrix}
\bar{\theta}_0 \\
\bar{\theta}_1
\end{pmatrix}, \Sigma.
$$

(24)

Although a further stage of modeling is still possible, we estimate the hyperparameters in (24) from data. In particular, we fit a first Fourier harmonic function to $\theta_0$ and $\theta_1$:

$$
\begin{aligned}
\bar{\theta}_0 &= \phi_{00} + \phi_{01} \sin \left( \frac{2\pi t}{12} \right) + \phi_{02} \cos \left( \frac{2\pi t}{12} \right) \\
\bar{\theta}_1 &= \phi_{10} + \phi_{11} \sin \left( \frac{2\pi t}{12} \right) + \phi_{12} \cos \left( \frac{2\pi t}{12} \right)
\end{aligned}
$$

(25, 26)

where

Fig. 10. Posterior simulations of monthly flow of nine sites not included previously in the MCMC simulation. The gray shaded region shows the 95% Bayesian interval. Black lines show observed data. The correlation between the expected flow from the Bayesian simulations and the observed flow is shown in the top left of the figure. Flow units are in m$^3$/s.
The resulting curve is shown in Fig. 9. The associated residual variance is used as an estimate for \( \Sigma \).

For the covariance matrix \( \Sigma \) in (22), one could also assume a Jeffreys prior, since the number of parameters (2) is small (Gelman et al., 2004). However, as is the case for the hierarchical Bayesian model for the annual maximum flow, the off diagonal elements of \( \Sigma \) and the variance of \( \theta_{ij} \) are relatively small, which in turn slow the MCMC convergence. Hence, we propose a conjugate prior for \( \Sigma \) with hyperparameters \( \lambda_i \) and \( \nu_i \) estimated from data:

\[
\Sigma_j \sim \text{Inv-Wishart}_{ij}(A_j).\tag{27}
\]

The following likelihood function (21) with the prior distributions (22), (23), (24), (27), yields the posterior distribution of the entire set \( \Theta \) of model parameters:

\[
p(\Theta|y) = \prod_{k=1}^{K} \prod_{i=1}^{n_i} \prod_{j=1}^{12} \mathcal{N}(\log(y_{ijk})|\theta_{0ij} + \theta_{ij}x(k), \sigma^2_j) 
\cdot \mathcal{N}(\theta_{0ij}|\theta_{0ij}, \Sigma_0) 
\cdot \mathcal{N}(\theta_{ij}|\theta_{ij}, \Sigma_j^{-1})
\cdot \text{Inv-Wishart}_{ij}(\Sigma_j|A_j).	ag{28}
\]

Similar procedure as done for (20) can be used here to obtain samples of parameters from the posterior distribution (28). Fig. 10 shows Bayesian posterior simulations (predictive check) of the monthly streamflow of nine randomly selected out-of-sample sites. For most sites, a good agreement between simulations and observed data is achieved, with correlations between the expected and observed streamflow up to 0.98. The proposed model is also able to reproduce the seasonality of the streamflow (Fig. 11) across the out-of-sample sites.

**Summary and discussion**

Our objective in this paper was to illustrate how hierarchical Bayesian models could be developed to provide a formal framework for estimating the uncertainty in hydrologic scaling relationships and the potential nonstationarity due to climate or other factors. The process was introduced by providing a framework...
for the Bayesian estimation of the scaling relationship with drainage area for annual maximum flows, using the Gumbel distribution as an example with Brazilian data. This model was then extended to illustrate how time varying scaling parameters could be estimated and used to assess whether there are statistically significant trends in the data. The model was then extended to consider monthly streamflow where in addition to the nonstationarity across years, one can account for the seasonal variation in the scaling relationships such that monthly flows at ungauged sites could be reproduced over the historical period. The obvious extension of the modeling framework introduced to the inclusion of climate predictors or trend informing variables and hence providing a basis for seasonal forecasting or climate change scenario development was not pursued. Work on this extension is under progress and will be communicated separately.

The results presented clearly demonstrate how regionalization of flow in terms of scaling parameters can be achieved in a non-stationary setting with uncertainty characterization and applied to the successful estimation of flow parameters at ungauged locations. We recognize that much of the effort in recent years is focused on the parameterization of highly detailed hydrologic water balance and stimulus–response models, and relatively little effort continues to be expended on empirical modeling of phenomena and uncertainty characterization. We see these as complementary approaches, and view our work as contributing directly to operational hydrology and water resources management by providing estimates of key quantities needed for design and management using the limited data sets that are typically available while recognizing the key issues that need to be addressed. The framework we present also presents a set of performance targets that more physically based methods need to be able to achieve if they are to be informative even given that they typically require dramatically higher amounts of data and processing. Similarly, in terms of process understanding, the framework presented here provides the capacity to explore specific hypotheses and parametric structures that an investigator may propose to explore the structure of the underlying processes, e.g., rainfall or other climate or land use data could be introduced as model predictors with appropriate functional forms for the predictive relationships. In the limit, a dynamical Bayesian network could be considered to model the dynamics of flow and information transfer across the physical network.

Acknowledgments

We would like to thank Vijay Gupta and the anonymous reviewer for the insightful comments that greatly improved the original manuscript.

References
