AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2016WR019676

Predicting nonstationary flood frequencies: Evidence supports an updated stationarity thesis in the United States

Adam Luke[1], Jasper A. Vrugt[1,2], Amir AghaKouchak[1], Richard Matthew[3], and Brett F. Sanders[1,3]

[1]Department of Civil and Environmental Engineering, University of California Irvine, Irvine, California, USA, [2]Department of Earth System Science, University of California Irvine, Irvine, California, USA, [3]Department of Planning, Policy, and Design, University of California Irvine, Irvine, California, USA

Key Points:
- Stationary predictions of flood peak distributions are preferred, overall
- Extrapolation of the nonstationary model parameter trend rarely improves the stationary prediction, even if an observed trend continues
- Using the most recent nonstationary parameters to predict with an updated stationary model is preferred for physically changing watersheds

Supporting Information:
- Supporting Information S1
- Supporting Information S2
- Supporting Information S3
- Supporting Information S4
- Supporting Information S5
- Supporting Information S6
- Supporting Information S7
- Supporting Information S8
- Supporting Information S9
- Data Set S1

Correspondence to:
A. Luke,
aluke1@uci.edu

Citation:
Luke, A., J. A. Vrugt, A. AghaKouchak, R. Matthew, and B. F. Sanders (2017), Predicting nonstationary flood frequencies: Evidence supports an updated stationarity thesis in the United States, *Water Resour. Res.*, *53*, 5469–5494, doi:10.1002/2016WR019676.

Received 24 AUG 2016
Accepted 26 NOV 2016
Published online 5 JUL 2017

Abstract Nonstationary extreme value analysis (NEVA) can improve the statistical representation of observed flood peak distributions compared to stationary (ST) analysis, but management of flood risk relies on predictions of out-of-sample distributions for which NEVA has not been comprehensively evaluated. In this study, we apply split-sample testing to 1250 annual maximum discharge records in the United States and compare the predictive capabilities of NEVA relative to ST extreme value analysis using a log-Pearson Type III (LPIII) distribution. The parameters of the LPIII distribution in the ST and nonstationary (NS) models are estimated from the first half of each record using Bayesian inference. The second half of each record is reserved to evaluate the predictions under the ST and NS models. The NS model is applied for prediction by (1) extrapolating the trend of the NS model parameters throughout the evaluation period and (2) using the NS model parameter values at the end of the fitting period to predict with an updated ST model (uST). Our analysis shows that the ST predictions are preferred, overall. NS model parameter extrapolation is rarely preferred. However, if fitting period discharges are influenced by physical changes in the watershed, for example from anthropogenic activity, the uST model is strongly preferred relative to ST and NS predictions. The uST model is therefore recommended for evaluation of current flood risk in watersheds that have undergone physical changes. Supporting information includes a MATLAB® program that estimates the (ST/NS/uST) LPIII parameters from annual peak discharge data through Bayesian inference.

1. Introduction

Dynamic flood risk is common: both natural and anthropogenic influences can alter flood behavior. For example, the rapid development of watersheds and physical alterations to rivers and coastlines contributed to variable flood risk throughout the twentieth century [*Villarini and Smith*, 2009; *Peel and Blöschl*, 2011; *Vogel et al*., 2011]. A warming climate increases the water holding capacity of the atmosphere which directly affects precipitation extremes and flood risk over time [*Karl et al*., 2009; *Trenberth*, 2011; *Cheng and AghaKouchak*, 2014]. Several studies have quantified societal exposure to current and intensified flooding in the 21st century [*Hallegatte et al*., 2013; *Hirabayashi et al*., 2013; *Milly et al*., 2002; *Apel et al*., 2006], and in the absence of adaptation strategies project a staggering global loss on the order of 1 trillion US dollars per year by 2050 [*Hallegatte et al*., 2013]. Indeed, the threat of escalating exposure requires new approaches for both flood risk management and characterization.

The most commonly applied method to characterize flood risk is known as flood frequency analysis (FFA), which estimates the recurrence rate of rare flooding events based on the annual exceedance probability of flood discharges. In the United States, engineers utilize FFA to determine the 100 year flood discharge, which is a design consideration for hydraulic structures and used to delineate areas subject to mandatory flood insurance. The 100 year discharge is expected to occur or be exceeded once every 100 years and has an estimated annual exceedance probability of 1%. In general, there are two commonly applied methods to estimate the exceedance probabilities necessary for FFA [*Centre for Ecology and Hydrology*, 2008]. In the first approach, a computer model simulates the rainfall-runoff relationship by numerical solution of the governing hydrologic equations. The calibrated model then estimates the flood discharges for different rainfall events with established exceedance probabilities. The second and alternative approach does not rely on a numerical model but uses only historic observations of flood discharges. This method is based on extreme

value analysis (EVA), where a probability distribution is fitted to the record of flood discharges themselves. The fitted distribution is then used to estimate the exceedance probabilities of specific flood discharges. The latter method is preferred in practice if sufficiently long records of flood discharges are available [*Centre for Ecology and Hydrology*, 2008] and is therefore the subject of the present study.

Both of these commonly applied methods are built on the assumption of constant flood frequency with respect to time, which implies the statistics of extreme discharges (e.g., mean and/or standard deviation) are time invariant. This so-called stationary (ST) assumption is quite restrictive, since a warming climate creates the possibility of unprecedented changes to observed flood regimes [*Min et al*., 2011; *Kunkel et al*., 2013], and anthropogenic watershed influences can alter the mean and variance of peak flood peaks [*Villarini and Smith*, 2009; *Vogel et al*., 2011]. The known limitations of ST FFA, along with the alarming societal impacts of intensified flooding, have given rise to the development and use of nonstationary (NS) approaches. This alternative FFA method does not rely on the ST assumption, but rather recognizes time variant flood frequencies.

Several different approaches have been published in the literature to handle the so-called "NS issue" in FFA [*Olsen*, 2002; *Raff et al*., 2009; *Jain and Lall*, 2001], and new national and state wide policies actually require the consideration of changing flood frequencies for planning, design, and risk management [*European Commission*, 2007; *Salas et al*., 2012]. An excellent review of such methods applied in Europe was reported by *Madsen et al*. [2013]. The two NS FFA methods described in this review include (1) the use of precipitation projections from future climate scenarios in rainfall-runoff models and (2) the use of a safety margin to adjust the design flood estimates derived from ST extreme value analysis (SEVA). While countries such as Norway, United Kingdom, Belgium, and Germany have adopted the safety margin approach in engineering guidelines [*Madsen et al*., 2013], the underlying ST assumption is often challenged and relaxed in the scientific literature to allow application of NS extreme value analysis (NEVA) [*AghaKouchak et al*., 2013; *Begueria and Vicente-Serrano*, 2006; *Begueria et al*., 2011; *Cheng et al*., 2014; *Cooley*, 2009; *Gilleland and Katz*, 2011; *Griffis and Stedinger*, 2007; *Katz*, 2010; *Lopez and Frances*, 2013; *Salas and Obeysekera*, 2014; *Silva et al*., 2015; *Stedinger and Griffis*, 2011; *Tramblay et al*., 2014; *Cheng et al*., 2015; *Villarini et al*., 2009; *Jakob*, 2013; *Steinschneider and Lall*, 2015]. Indeed, the review by *Madsen et al*. [2013] concludes on page 33 that moving toward a new NS framework, based on the use of NEVA, is an "important aspiration within the European hydrological science community."

NEVA is an important extension of SEVA which enables the parameters of the extreme value distribution to vary with time. This allows the parameters of the distribution to capture trends in flood frequencies [*Salas and Obeysekera*, 2014]. In theory, these trends can be extrapolated to estimate, for example, the distribution of flood discharges in 2050. Numerous studies have shown that NEVA improves the statistical representation of historic hydroclimatic data [*Lopez and Frances*, 2013; *Strupczewski et al*., 2001; *Villarini et al*., 2009], but the authors consistently caution against using NEVA for out-of-sample prediction. In this study, we apply split-sample testing to 1250 annual maximum discharge records and compare the predictive capabilities of NEVA relative to SEVA using a log-Pearson Type III (LPIII) distribution. Specifically, we reserve the second half of each data record to evaluate predictions under the calibrated ST and NS models. We use two different approaches to predict the out-of-sample data with the NS model, which are described along with the ST/NS models in section 2. Section 3 includes a description of the United States Geological Survey (USGS) records used in this study. The parameter estimation procedure is presented in sections 4.1–4.1.3, where the parameters of the LPIII distribution in the ST and NS models are inferred from the first half of each record using Markov Chain Monte Carlo (MCMC) simulation with the DREAM$_{(ZS)}$ algorithm. The posterior parameter distributions and model fit metrics are compared in section 4.1.4, along with a discussion of issues associated with NS model section. Section 4.2 outlines our application of Bayesian hypothesis testing which we use to evaluate, compare, contrast, and juxtapose the predictions made by the competing models. The results of this analysis are presented in sections 5 and 6, where we assess (1) which of the three predictions, under the LPIII distribution, describe most accurately the distribution of the out-of-sample flood discharges and (2) which diagnostics are useful for model selection. The paper concludes in section 7 with recommended applications of NEVA for prediction of flood discharges in the LPIII distribution. Appendix A includes the mathematical definition of the LPIII distribution, and the supporting information includes a MATLAB® program based on the methods outlined in sections 4.1–4.1.3 which can be used for inference of the ST/NS LPIII model parameters.

2. Stationary or Nonstationary?

FFA in the United States is conducted according to the guidelines outlined in Hydrologic Bulletin 17B [*Interagency Committee on Water Data*, 1982], which are built on the ST assumption. NS extensions of the standardized FFA guidelines in the United States have been considered in several studies [*Stedinger and Griffis*, 2011; *Griffis and Stedinger*, 2007]. For instance, *Stedinger and Griffis* [2011] use a time-dependent mean to account for NS in the Pearson Type III (PIII) distribution used widely for FFA in the United States and Australia [*Pilgrim*, 2001]

$$\log_{10}(Q) \sim \text{PIII}(\mu_t, \sigma, \gamma) \tag{1}$$

where Q signifies the annual maximum discharge, t denotes time, and μ_t, σ, and γ characterize the mean, standard deviation, and skewness of the PIII distribution, respectively. The $\log_{10}$-transformation stabilizes the variance of the annual peak discharges and makes the transformed data more amenable to the PIII distribution. We can assume a simple linear trend for the mean of the PIII distribution

$$\mu_t = \mu_o + \alpha t \tag{2}$$

where μ_o signifies the intercept and α is the slope of the (t, μ_t) relationship. The value of μ_o is equivalent to the PIII mean of the first year of the discharge record. If flood peaks are assumed to be stationary then $\alpha = 0$ suffices in equation (2) and the resulting model simplifies to the ST PIII distribution. Joint Bayesian inference of the PIII parameters $\{\sigma, \gamma\}$ and latent variable α from a n-record, $\mathbf{Q}=\{Q_1, \ldots, Q_n\}$, of observed annual peak discharges, would help judge which statistical model, NS or ST, would receive most support from the data. This would require knowledge of the marginal likelihood, or model evidence, which integrates model accuracy, uncertainty, and complexity. Appendix A lists analytic expressions for the probability density and cumulative distribution functions of the PIII distribution. We adopt the notation LPIII to clarify the use of log-transformed discharge data, and refer to equation (1) as ST LPIII model ($\alpha = 0$) or NS LPIII model ($\alpha > 0$). For clarity, we use μ to denote the time invariant mean of $\log_{10}(\mathbf{Q})$ in the ST LPIII model.

The use of a linear time-dependency of the LPIII mean in equation (2) is rather convenient and simplistic, but has several important implications. First, a trend in the logarithmic mean of Q equates to an exponential trend in the mean of the flood peaks [*Read and Vogel*, 2016]. This assumption may seem questionable but is supported by analysis of flood trends in the United States [*Vogel et al.*, 2011] and United Kingdom [*Prosdocimi et al.*, 2014]. Second, despite the use of a constant value for σ, the standard deviation of the arithmetic flood peaks will increase with the logarithmic mean, μ_t, of the LPIII distribution. This intrinsic property is desirable since nonstationarity would likely not only alter the mean of the flood peaks but simultaneously also affect their associated dispersion. Third, equation (2) does not provide guidance on the governing physical factors, climate signals, and/or anthropogenic causes that may explain temporal changes in the frequencies of the annual flood peaks. Fortunately, the parameter estimation methodology presented herein supports the use of much more advanced (t, μ_t) relationships with covariates other than time. Examples include changes in land-use, urbanization, and variations in the North Atlantic Oscillation (NAO) and El Nino Southern Oscillation (ENSO). Yet, these covariates are not always readily available, nor may they correlate sufficiently with flooding events [*Archfield et al.*, 2016]. We, therefore, resort herein to a purely statistical description of annual discharge peaks and use time as a proxy and explanatory variable of nonstationarity. In fact, the LPIII model used herein (with $\alpha > 0$) is analogous to the simplest of NS models presented by *Stedinger and Griffis* [2011].

Figure 1 demonstrates the additional flexibility attained through NEVA relative to SEVA. Figure 1a shows the effect of γ on the ST LPIII model, while Figure 1b shows the effect of α on the NS LPIII model. Notice that the traditional ST LPIII model represents different tailing behavior as γ changes, but the distribution must remain constant in time. Indeed, the NS LPIII model includes only one additional parameter, α, but adds the dimension of time to our EVA. However, it is not clear if predictions of flood peak distributions under a NS model are an improvement relative to ST predictions, despite the additional flexibility. The primary concern is whether the NS signal is clear enough to accurately estimate trend parameters, such as α in this study. Moreover, different NS models will lead to different predictions, and there are multiple approaches for prediction under a single NS model. This raises the question of whether we actually do better by including NS in our analysis [*Stedinger and Griffis*, 2011].

To illustrate these issues, Figure 2a shows the ST and NS LPIII models fit to the first half of the annual maximum discharge record from the Smith River near Crescent City, CA (for details see section 4). Here we use

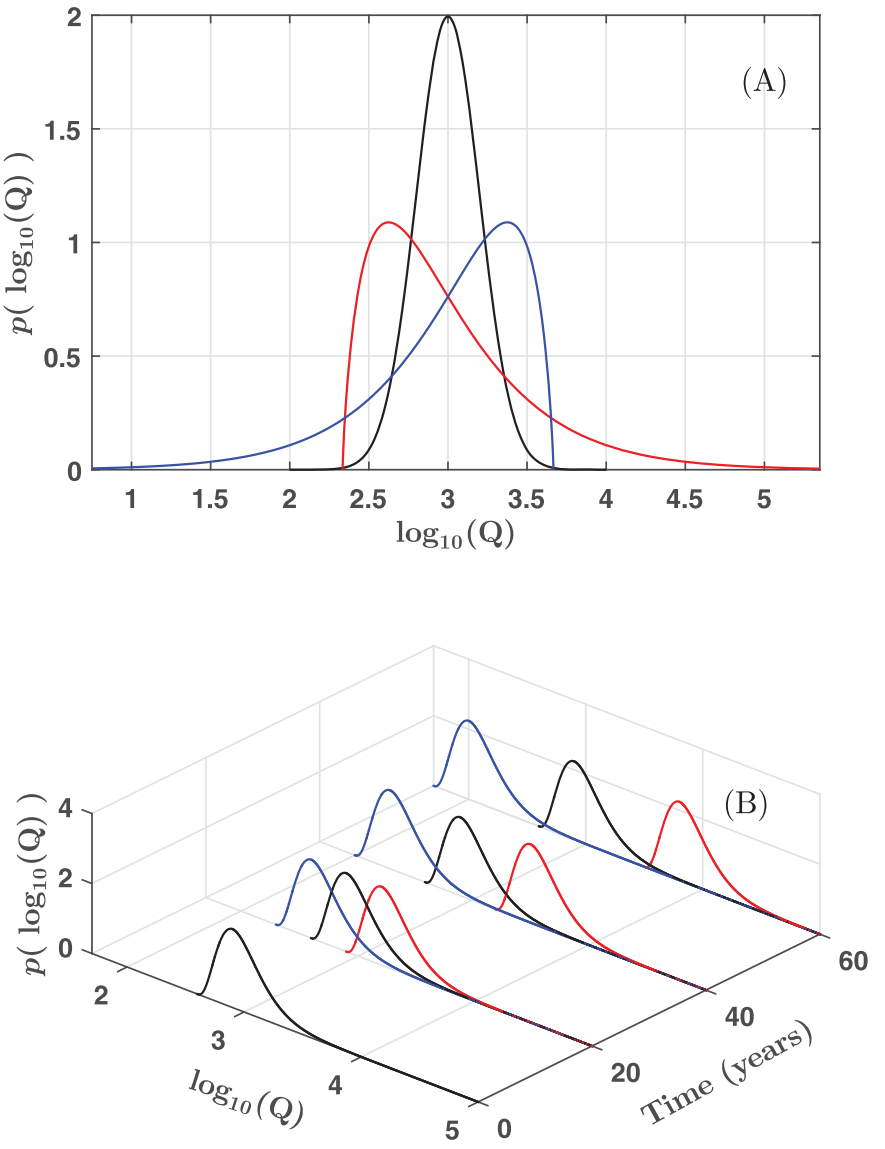

Figure 1. (A) ST LPIII distribution with variable shape parameter. The *red*, *black*, and *blue* lines correspond to γ values of 1.5, 0, and −1.5, respectively. (B) NS LPIII distribution with variable trend parameter. The *red*, *black*, and *blue* lines correspond to α values of 0.015, 0, and −0.015, respectively.

two different approaches to predict the out-of-sample data with the NS model. Under the first approach, the trend of the NS model parameters is extrapolated throughout the evaluation period. In the second approach, the NS model parameter values at the end of the fitting period are used to predict with a ST model, or the ST parameters are "updated" by the NS model (hereafter denoted uST). Including the ST LPIII model leads to three different representations of the out-of-sample density, which are shown relative to the in-sample density in Figures 2b–2d. During the fitting period, the annual maximum discharge of the Smith River exhibits a detectable trend at the 0.05 significance level (Mann-Kendall trend test) [*Mann*, 1945; *Kendall*, 1976], which is reflected by the trend in μ_t. In the NS case, μ_t has substantially increased by the end of the fitting period, indicating that the mean and scale of the annual maximum discharges have changed throughout the record. In the ST case, μ remains constant. The difference in μ between the two approaches is certainly significant for planning purposes and engineering design. The 99th percentile of the ST LPIII distribution (q_{99}) is used to designate the special flood hazard area and design critical infrastructure such as bridges and levees in the United States [*Federal Emergency Management Agency*, 2009; *Interagency Committee on Water Data*, 1982]. Under the NS LPIII model at the end of the fitting period (or uST model), q_{99} is

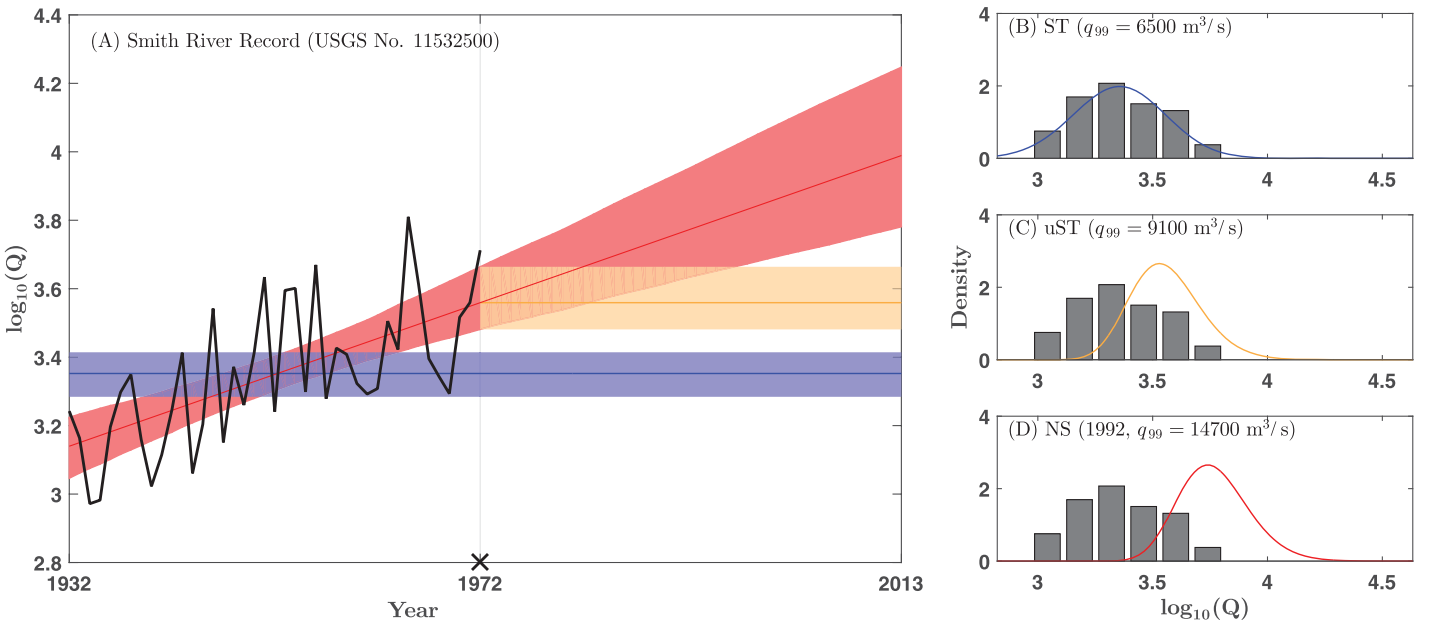

Figure 2. (a) Maximum a posteriori (MAP) estimate of the LPIII mean under the ST (*blue* line), uST (*gold* line), and NS (*red* line) models inferred from the Smith River fitting period **Q** (*black* line). The colored shading represents the respective 95% credible intervals of the LPIII mean, and the *black* cross denotes the end of the fitting period. (b–d) Predictions of out-of-sample density under the ST (*blue* line), uST (*gold* line), and NS (*red* line) models derived from the MAP parameter estimates. The *black* histograms represent the empirical density of the fitting period. Notice that Figures 2c and 2d show predictions under the uST and NS models moving *away* from the observed density, and the 95% credible intervals are wider under the NS and uST models relative to the ST model.

equivalent to 9100 m^3/s, which is 40% greater than the ST estimate (Figures 2b–2d). Assuming that flood frequencies are indeed nonstationary, and the distribution of annual maximum discharges continues to change according to the best fit NS model parameters, the value of q_{99} under the NS model is 14,700 m^3/s by the year 1992 and 2.3 times larger than its counterpart of the ST model.

The application of the NS LPIII model to stationary discharge records can lead to overestimation of flood peaks, and hence overengineered structures. Even worse, falsely selecting the ST model could lead to underdesigned infrastructure and potential disaster. In the absence of formal guidelines on NS model selection and lack of recommendations for FFA with a trend in the historic record, it would be difficult to judge which model to use in practice. Moreover, it is unclear how to use a NS model for prediction. Is it better to assume flood frequencies will continue to change according the trends inferred from the historic record, or should we simply use the most recent NS parameterization for prediction? These issues motivate the split-sample testing used herein, where the predictions of the calibrated ST, uST, and NS models are evaluated using Bayesian hypothesis testing.

3. Stream Gage Records

We use a large data set of annual maximum discharge records to evaluate the predictive abilities of the ST, uST, and NS models. The majority of records used in this study originate from the set of 1861 stream gages which were selected for potential use in the Model Parameter Estimation Project (MOPEX) [*Schaake et al.*, 2006; *Slack and Landwehr*, 1992; *Wallis et al.*, 1991]. The gages were believed to be unaffected by upstream regulation and watershed development at the time the data set was compiled (1991/1992), suiting the needs of MOPEX. Since then, many records experienced anthropogenic changes. We rely on the USGS data flags "5", "6" and "C" to identify records affected by upstream regulation and/or diversion (flags "5" and "6"), or land-use changes and/or channelization (flag "C"). In this study, watersheds that have experienced anthropogenic influences are included in the analysis since the ST assumption is likely violated under these circumstances. An additional 25 watersheds affected by land use changes or channelization were added to the record pool since relatively few gages of this category were included in the original MOPEX data set. Records with fewer than 60 observations were excluded to ensure the availability of at least 30 observations in the fitting and evaluation period.

We denote the full record of measured annual maximum discharges, **Q**, and the fitting period data, **X**. The fitting period data include the first half of measured peak discharges, $\mathbf{X}=\log_{10}(\{Q_1,\ldots,Q_n\})$, where Q denotes an observation within the full record, $n=\lceil N/2 \rceil$, and N is the number of observations in the full record. The evaluation data, $\mathbf{X}^*$, includes the remaining observations, or $\mathbf{X}^*=\log_{10}(\{Q_{n+1},\ldots,Q_N\})$. Split sampling in this manner resulted in variable fitting period lengths, with 40 observations representing the average data availability for parameter inference. In this study, historic peaks were removed from the record. Also, records with no-flow observations (i.e., zero discharge) were excluded from the analysis to simplify the use of the LPIII distribution. This leads to a data set of 1250 annual maximum discharge records. Figure 3a displays the geographic location of the different records, while Figures 3b–3d present various characteristics of these basins and their fitting period discharge records. Trend analysis shown in Figure 3 was conducted according to the Mann-Kendall test for monotonic trends [*Kendall*, 1976; *Mann*, 1945]. Hereafter, a "significant trend" refers to a rejection of the null hypothesis of no trend by the Mann-Kendall trend test at the 0.05 significance level. We acknowledge that there are no good statistical reasons to omit historic peaks and records with zero flow observations from the data set [*Reis and Stedinger*, 2005; *Interagency Committee on Water Data*, 1982], yet the focus of this study is on systematically assessing predictions made under ST, uST, and NS models.

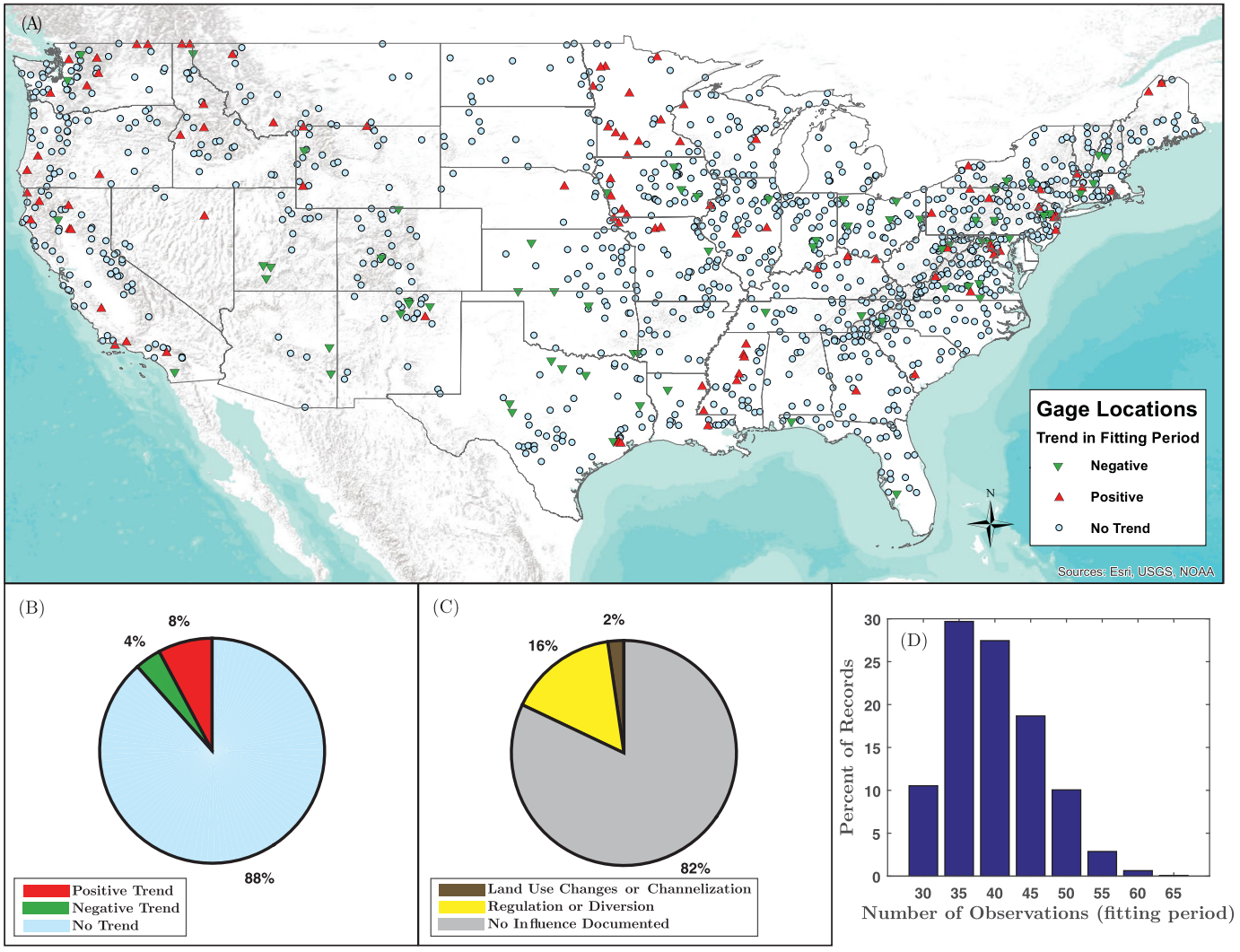

Figure 3. (a) Location of annual maximum discharge records and trends detected according to the Mann-Kendall trend test at the 0.05 significance level. (b) Fitting period trend characteristics as a percentage of all records tested. (c) Anthropogenic watershed influences as a percentage of all records tested. Land use changes or channelization refers to USGS record flag "C," while regulation and diversion refers to USGS record flag 5 or 6. (d) Fitting period length as a percentage of all records tested.

4. Bayesian Inference

4.1. Parameter Estimation

Estimating the parameters of an extreme value distribution is important for reliable frequency analysis. This process is further complicated by the introduction of NS models, where structural-trend parameters must also be inferred from limited data. In this study, we are primarily interested in inferring the ST LPIII parameters, $\boldsymbol{\theta}_s=\{\mu,\sigma,\gamma\}$, and the NS LPIII parameters, $\boldsymbol{\theta}_n=\{\mu_o,\sigma,\gamma,\alpha\}$, where $\mu_t=\mu_o+\alpha t$. While there are certainly other viable parameterizations of the NS LPIII model, we do not consider NS models describing changes in σ and γ because it is unlikely that these trends can be estimated from common record lengths [*Yu et al.*, 2015]. We also do not consider trend models more complex than the simple linear model, due to the uncertainty introduced by additional trend parameters. Uncertainty introduced by NS model complexity has been reported as a major drawback to NEVA, which is the primary argument against widespread application made by *Serinaldi and Kilsby* [2015]. Therefore, any meaningful comparison between flood discharge predictions based on ST and NS models must account for parameter uncertainty introduced by the NS model. The issue of parameter uncertainty motivates the Bayesian approach for parameter inference and model comparison. Bayesian parameter estimation results in distributions of parameter values rather than point estimates, which is especially useful for hypothesis testing and uncertainty quantification. Thus, we adopt Bayesian inference and use MCMC simulation with the DREAM$_{(ZS)}$ algorithm to robustly estimate the posterior distribution of the NS/ST/uST LPIII parameters and quantify model predictive uncertainty.

Under the Bayesian approach for parameter estimation, Bayes' theorem is used to calculate the posterior density of parameter values

$$p(\boldsymbol{\theta}_j|\mathbf{X},\mathcal{M}_j)=\frac{p(\boldsymbol{\theta}_j|\mathcal{M}_j)L(\boldsymbol{\theta}_j|\mathbf{X},\mathcal{M}_j)}{p(\mathbf{X}|\mathcal{M}_j)} \tag{3}$$

where $p(\boldsymbol{\theta}_j|\mathbf{X},\mathcal{M}_j)$ represents the posterior density of parameter vector, $\boldsymbol{\theta}_j$, given the fitting period data, $\mathbf{X}$, and the model of interest, $\mathcal{M}_j$. The subscript j denotes competing model classes, and in this study, $j=\{$"s," "n," "u"$\}$, which stand for the ST, NS, and uST models, respectively. Here $p(\boldsymbol{\theta}_j|\mathcal{M}_j)$ represents the prior density of $\boldsymbol{\theta}_j$ given $\mathcal{M}_j$, and $L(\boldsymbol{\theta}_j|\mathbf{X},\mathcal{M}_j)$ represents the likelihood of parameter values $\boldsymbol{\theta}_j$, given $\mathbf{X}$ and $\mathcal{M}_j$. The denominator, $p(\mathbf{X}|\mathcal{M}_j)$, is known as the marginal likelihood or evidence and acts as a normalization constant so that $p(\boldsymbol{\theta}_j|\mathbf{X},\mathcal{M}_j)$ integrates to one. The evidence is obtained through integration over the parameter space

$$p(\mathbf{X}|\mathcal{M}_j)=\int p(\boldsymbol{\theta}_j|\mathcal{M}_j)L(\boldsymbol{\theta}_j|\mathbf{X},\mathcal{M}_j)d\boldsymbol{\theta}_j \tag{4}$$

where $p(\mathbf{X}|\mathcal{M}_j)$ also quantifies the probability of seeing the data that were actually observed under the competing models. Therefore, the evidence is particularly useful for model selection, because the model which maximizes equation (4) is a better descriptor of the data. Bayesian hypothesis testing and the evidence will be discussed further in section 4.2. However, Bayesian parameter estimation does not necessarily require the value of $p(\mathbf{X}|\mathcal{M}_j)$, since parameter inference can be made from the unnormalized density

$$p(\boldsymbol{\theta}_j|\mathbf{X},\mathcal{M}_j)\propto p(\boldsymbol{\theta}_j|\mathcal{M}_j)L(\boldsymbol{\theta}_j|\mathbf{X},\mathcal{M}_j) \tag{5}$$

when MCMC methods are used to sample from the posterior, as applied herein. For now, we are interested in using equation (5) to estimate the unnormalized posterior density of both $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_n$ given the first half of annual maximum discharge records. Equation (5) is not directly applied to infer $\boldsymbol{\theta}_u$, since the parameters of the uST model are derived from $p(\boldsymbol{\theta}_n|\mathbf{X},\mathcal{M}_n)$. In most practical applications, estimating $p(\boldsymbol{\theta}_j|\mathbf{X},\mathcal{M}_j)$ is accomplished by sampling from the posterior after specifying $p(\boldsymbol{\theta}_j|\mathcal{M}_j)$ and $L(\boldsymbol{\theta}_j|\mathbf{X},\mathcal{M}_j)$.

4.1.1. Likelihood Function (Fitting Period)

Conceptually, $L(\boldsymbol{\theta}_j|\mathbf{X},\mathcal{M}_j)$ represents how plausible a parameter combination is after observing $\mathbf{X}$, which quantifies the information content of the data. Specifically, $L(\boldsymbol{\theta}_j|\mathbf{X},\mathcal{M}_j)$ is the probability density of $\mathbf{X}$ according to $\mathcal{M}_j$ with parameterization $\boldsymbol{\theta}_j$. For the parameter inference procedure, the likelihood function is defined as

$$L(\boldsymbol{\theta}_j|\mathbf{X},\mathcal{M}_j)=\prod_{i=1}^{n}p(X_i|\boldsymbol{\theta}_j,\mathcal{M}_j) \tag{6}$$

where n is the number of observations in the fitting period, and X_i represents the ith observation. In this study, $p(X_i|\boldsymbol{\theta}_j,\mathcal{M}_j)$ is obtained by evaluating an analytic expression of the PIII probability density function (equation (A1) or (A3)) at observation $X_i=\log_{10}(Q_i)$ and parameterization $\boldsymbol{\theta}_j$ under $\mathcal{M}_j$. Here $j=\{$"s," "n"$\}$,

which distinguishes equation (6) applied to determine the likelihood of the ST and NS model parameters, respectively.

Under $\mathcal{M}_s$, the parameter values within $\boldsymbol{\theta}_s = \{\mu, \sigma, \gamma\}$ are constant with respect to time, and therefore equation (6) is evaluated at a fixed parameterization of the PIII density function for $i = \{1, \ldots, n\}$. However, this is not the case under $\mathcal{M}_n$. When equation (6) is applied to calculate the likelihood of $\boldsymbol{\theta}_n = \{\mu_o, \sigma, \gamma, \alpha\}$, the parameterization of the PIII density functions changes with $i = \{1, \ldots, n\}$, according to the linear trend model. Therefore, at each observation X_i, there is a unique parameterization of the PIII density function

$$\boldsymbol{\theta}_n^i = \{\mu_o + \alpha t_i, \sigma, \gamma\} \tag{7}$$

where t_i is the time (in years) between the first and the ith observation. In this study, t_i is considered the covariate, since $\mathcal{M}_n$ parameter values depend on, and change with, t_i. For a more general description of covariate modeling, see *Renard et al*. [2013] and *Coles* [2001]. There are several properties of $\mathcal{M}_s$ and $\mathcal{M}_n$ with respect to the likelihood function that are worth noting.

First, $\mathcal{M}_s$ is a special case of $\mathcal{M}_n$ when $\boldsymbol{\theta}_n^i$ is constant or $\alpha = 0$. For records without a strong trend, equation (6) will be maximized near $\alpha = 0$ under $\mathcal{M}_n$, and therefore the parameter values of $\boldsymbol{\theta}_n$ will approach $\boldsymbol{\theta}_s$. In this scenario, α becomes a nuisance parameter, and $\mathcal{M}_n$ is overparameterized. This suggests that if α is inferred using maximum likelihood, there must be a time-dependent trend in $\mathbf{X}$ for $\mathcal{M}_n$ to significantly differ from $\mathcal{M}_s$. Second, assuming that a trend is observed in $\mathbf{X}$, $\mathcal{M}_n$ has a much greater ability to maximize equation (6) relative to $\mathcal{M}_s$. This is because in the NS case, the location of the distribution changes with the trend in $\mathbf{X}$, enhancing the probability density near the observations relative to the fixed μ under $\mathcal{M}_s$. The NS model's ability to maximize equation (6) relative to the ST model is problematic for comparison of $\mathcal{M}_n$ and $\mathcal{M}_s$ based on likelihood metrics alone, since likelihood based metrics will almost certainly favor $\mathcal{M}_n$ for records with detected trends.

4.1.2. Prior Distributions (Fitting Period)

While $L(\boldsymbol{\theta}_j|\mathbf{X}, \mathcal{M}_j)$ represents the plausibility of $\boldsymbol{\theta}_j$ after considering the data, $p(\boldsymbol{\theta}_j|\mathcal{M}_j)$ represents knowledge of $\boldsymbol{\theta}_j$ before the data are considered. Often, $p(\boldsymbol{\theta}_j|\mathcal{M}_j)$ is referred to as the prior. The prior represents an analyst's knowledge of $\boldsymbol{\theta}_j$ before any data are collected, which is described mathematically by the joint probability of the parameters within $\boldsymbol{\theta}_j$. The formal inclusion of prior knowledge about $\boldsymbol{\theta}_j$ in the parameter inference process is unique to Bayesian inference. The choice of priors is subjective, but it allows the analyst to use information other than the data for parameter inference.

While several studies have specified highly informative joint-priors for hydrologic data through regional analysis or eliciting expert opinion [*Behrens et al*., 2004; *Perreault et al*., 2000; *Renard et al*., 2006a], we decided to employ relatively uninformative priors so that inferred values of $\boldsymbol{\theta}_s$ are similar to parameterizations produced following standard practice in the United States. In the United States, μ and σ are determined entirely through consideration of $\mathbf{X}$, while γ is based on a regional and site specific estimate [*Interagency Committee on Water Data*, 1982]. An analogous parameterization can be produced in a Bayesian framework by using uninformative priors on μ and σ with an informative prior specified on γ.

Table 1. Model Parameters and Prior Distributions[a]

	Scale	Units	Prior	a	b
$\mathcal{M}_n$ Parameters					
μ_o	$\log_{10}$	m^3/s	$\mathcal{U}(a, b)$	−10	10
σ	$\log_{10}$	m^3/s	$\mathcal{U}(a, b)$	0	2
γ	$\log_{10}$		$\mathcal{N}(\gamma_r, SD_{\gamma_r})$		
α	$\log_{10}$	$m^3/(s\ yr)$	$\mathcal{U}(a, b)$	−0.15	0.15
$\mathcal{M}_s$ Parameters					
μ	$\log_{10}$	m^3/s	$\mathcal{U}(a, b)$	−10	10
σ	$\log_{10}$	m^3/s	$\mathcal{U}(a, b)$	0	2
γ	$\log_{10}$		$\mathcal{N}(\gamma_r, SD_{\gamma_r})$		

[a]The regional estimate of γ and γ_r was taken from the Bulletin 17B regional skew map for each record location, and SD_{γ_r} was set to 0.55 per Bulletin 17B guidelines. The multivariate prior density is simply the product of the marginal distributions evaluated at a proposed parameter combination.

For uninformative, yet proper priors, the marginal priors for μ, μ_o, σ, and α were specified as a uniform distribution, $\mathcal{U}(a, b)$, where a and b represent the lower and upper bounds of the uniform priors, respectively. Our primary reason for using bounded priors is algorithmic efficiency. Indeed, such prior enhances considerably the convergence speed of the $DREAM_{(ZS)}$ algorithm (section 4.1.3), a particularly important consideration in the face of the relatively large number of data records used in the present study. Table 1 lists the selected bounds on the uniform priors, which were chosen based on physical reasoning. For example, the bounds

on μ restrict the location of the distribution to values between 10^{-10} and 10^{10} m^3/s, which adequately encompass realistic values of μ. We confirmed that none of the posteriors were truncated by the prior bounds following the fitting procedure.

Selecting a distribution that represents a lack of knowledge is challenging and subjective, and the bounded uniform distribution is certainly not the only viable uninformative prior. For example, *Reis and Stedinger* [2005] use Jeffrey's prior as an uninformative prior on the scale parameter, and a normal a distribution with a large variance for an uninformative prior describing the location parameter. Both approaches will lead to inference of the location and scale parameters based on the likelihood of **X**, or the information content of the data alone. In contrast, we use an informative prior on γ because information other than **X** is typically available and used to estimate the skewness of annual maximum discharge records.

Since γ controls the tail behavior and extrapolation to the unobserved percentiles of the LPIII distribution, it is difficult to determine from a single record. Therefore, estimates of γ are usually based on a site specific and regional approximation. Bulletin 17B approximates γ by weighting the station estimate, γ_s, with a regional estimate, γ_r. In the Bulletin 17B guidelines, γ_s is determined by calculating the sample skew of **X**, and γ_r is typically taken from an isocline map of regional skews. The weights on γ_s and γ_r are inversely proportional to their respective mean square errors. A Bayesian adaptation of this weighting process can be accomplished for $\mathcal{M}_n$ and $\mathcal{M}_s$ through an informative marginal prior

$$p(\gamma|\mathcal{M}_j) \sim \mathcal{N}(\gamma_r, \mathrm{SD}_{\gamma_r}) \tag{8}$$

where SD_{γ_r} is the root-mean-square error of the regional skew, γ_r, and $\mathcal{N}$ denotes the normal distribution. Values of γ close to the regional estimate are assigned a high prior probability, and thus the prior reflects regional knowledge of γ. In this study, γ_r was obtained for each record by extracting values of γ_r from a digitized version of the Bulletin 17B Plate 1 generalized skew map. Here SD_{γ_r} was set equal to 0.55 per Bulletin 17B guidelines. Without considering the regional skews through the informative prior, posterior γ would be fairly different from γ estimated in practice. We also note that using an informative prior for only the shape parameter has been recommended for NS model parameter estimation in the generalized extreme value distribution (GEV) [*El Adlouni et al*., 2007], and this approach was previously designated the generalized maximum likelihood (GML) method in the ST case [*Martins and Stedinger*, 2000]. *Ouarda and El-Adlouni* [2011] apply the GML method for inference of NS GEV parameters in a Bayesian framework, which parallel the methods applied in this study. The difference herein is that the generalized prior on γ has been adapted to incorporate regional information.

4.1.3. MCMC Simulation With the DREAM$_{(ZS)}$ Algorithm

Now that $p(\boldsymbol{\theta}_j|\mathcal{M}_j)$ and $L(\boldsymbol{\theta}_j|\mathbf{X}, \mathcal{M}_j)$ have been defined, the posterior distribution of $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_n$ can be approximated using MCMC simulation with the DREAM$_{(ZS)}$ algorithm [*Vrugt et al*., 2009; *Laloy and Vrugt*, 2012; *Vrugt*, 2016]. Sampling methods such as MCMC are required for most applications of Bayes' theorem, since $p(\boldsymbol{\theta}_j|\mathbf{X}, \mathcal{M}_j)$ is often analytically intractable. Essentially, MCMC methods provide a sample from the posterior distribution, and parameter inference is made from the generated sample. A thorough explanation of Monte Carlo simulation appears in *Vrugt* [2016, section 2], and interested readers are referred to this publication for further details.

We use three different Markov chains to generate target samples and monitor convergence of the sampled chains using the $\hat{R}$-statistic of *Gelman and Rubin* [1992]. Specifically, for each model and record, *8,000* samples were created in each chain with a burn-in of 50%. This equates to a total of 12,000 realizations of $\boldsymbol{\theta}_j$ drawn from the target distribution. Parameter inference and model comparison is now based on the posterior samples of $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_n$ produced by the MCMC algorithm. The parameters of the uST distribution, $\boldsymbol{\theta}_u = \{\mu_n, \sigma, \gamma\}$, are derived from the posterior sample of $\boldsymbol{\theta}_n$. Here μ_n represents the parameter value of μ_t at the end fitting period under $\mathcal{M}_n$. The distribution of μ_n was obtained by applying equation (2) at each posterior sample of $\boldsymbol{\theta}_n$, with t equal to the time between the first and the last year of the fitting period. The uST σ and γ parameters were taken from the posterior sample of $\boldsymbol{\theta}_n$, and not $\boldsymbol{\theta}_s$.

4.1.4. Posterior Comparisons

We now return our attention to the Smith River data record, previously shown in Figure 2. Figure 4 presents the marginal posterior distributions of $\boldsymbol{\theta}_n$ (top row), $\boldsymbol{\theta}_s$ (middle row), and $\boldsymbol{\theta}_u$ (bottom row) given the data in the Smith River fitting period. The prior density and maximum likelihood parameter estimates are represented by the black lines and black crosses, respectively. The maximum likelihood estimate (MLE) of $\boldsymbol{\theta}_j$,

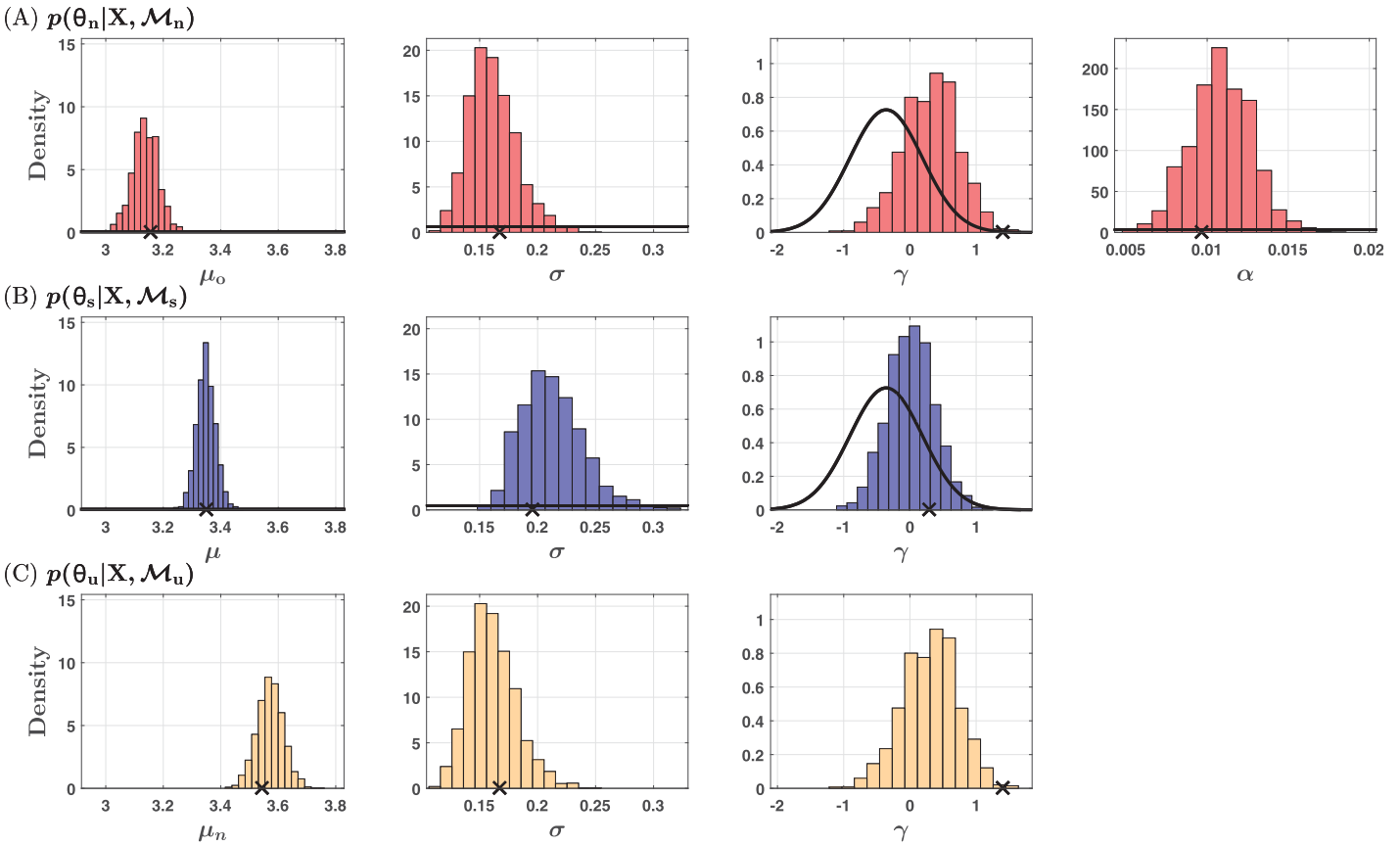

Figure 4. Marginal posterior distributions of θ_n under the NS model (*red* histograms), θ_s under the ST model (*blue* histograms), and θ_u under the uST model (*gold* histograms), given the data in the Smith River fitting period. *Black* lines represent the marginal prior distributions, and the *black* crosses show the MLE of θ_j under the three models. Notice that the MLE of γ under $\mathcal{M}_n$ is very unlikely according to the prior distribution, and posterior θ_n favors lower values of σ than posterior θ_s.

denoted $\hat{\theta}_j$, is taken as the parameter combination from the posterior sample which maximized equation (6). This figure highlights several important results. First, Figure 4 shows that the informative prior on γ causes the marginal posterior distribution to be centered between the MLE and the mode of the informative prior. Since the MLE of γ is the site specific skew, and the prior mode is the regional skew, posterior γ represents a combination of the site specific and regional information. This result demonstrates the utility of informative priors for the inference of shape parameters.

Second, notice that the posterior under $\mathcal{M}_n$ favors lower values of σ than the posterior under $\mathcal{M}_s$. This difference can be explained by how the two models represent records that exhibit a strong trend. The apparent trend in the data, which is shown in Figure 2, is simply variance about the mean of the data in the context of a ST model. Therefore, the observable difference in the magnitude of the annual maximum discharges from the beginning to the end of the fitting period is accounted for through a relatively large value of σ. However, under $\mathcal{M}_n$, the trend causes μ_t to increase throughout the fitting period. Thus, the time-dependent increase in the observed discharges is represented by a changing location parameter, not a large variance. In the context of $\mathcal{M}_n$, σ does not describe the variance of the entire sample, but rather the variance about the changing mean. This concept is further illustrated in Figures 1b–1d, where the ST distribution exhibits a variance large enough to encompass all of the data, but the NS distributions do not. The apparent difference in σ between the $\mathcal{M}_n$ and $\mathcal{M}_s$ has implications for inference of γ.

Third, let us compare the marginal posteriors of γ between θ_s and θ_n. While the posterior distribution of γ is relatively similar between the two models, the marginal distribution under $\mathcal{M}_n$ includes more density away from the prior mode. This is due to the difference in the value of the MLE between the two models. Under $\mathcal{M}_s$, the MLE of γ is closer to the regional estimate relative to $\mathcal{M}_n$. Indeed, the MLE of γ under $\mathcal{M}_n$ is found in the tail of the prior distribution. In other words, the MLE is very unlikely according to our prior knowledge.

As a consequence, the NS posterior is centered further to the right. This suggests that the likelihood function under $\mathcal{M}_n$ favors values of γ that are unrealistic according to past studies, which can be explained by the relatively low value of σ. Previous studies have also shown that σ and γ are inversely related in the LPIII distribution [*Srikanthan and McMahon*, 1981; *Nozdryn-Plotnicki and Watt*, 1979], so in general low σ estimates will produce large skews. Let us explore this further by comparing the MLE of γ under $\mathcal{M}_n$ and $\mathcal{M}_s$ for other records tested.

Figure 5a shows the empirical density of $\hat{\gamma}$ under $\mathcal{M}_n$ (red histogram) and $\mathcal{M}_s$ (blue histogram) for estimates obtained from fitting periods with relatively large trends. Large trends are defined as $\hat{\alpha}$ values one standard deviation away from the mean value of zero, which were found for 236 fitting periods. Values shown in Figure 5 were produced by the procedure outlined in section 4.1, except here the prior on γ was completely uninformative, or $p(\gamma|\mathcal{M}_j) \sim \mathcal{U}(-5, 5)$. This ensures that $\hat{\gamma}$ is adequately approximated, since the MCMC algorithm is no longer constrained by the regional estimate. We show how $\hat{\gamma}$ is distributed for records with large trends because the difference in the estimates under the two models generally increase with increasing values of $\hat{\alpha}$. Figure 5b also compares the empirical cumulative distribution function (CDF) of $\hat{\gamma}$ under $\mathcal{M}_n$ (red line) and $\mathcal{M}_s$ (blue line) and reveals that there is a greater probability of large $\hat{\gamma}$ values (i.e., ± 1.4) when estimated under $\mathcal{M}_n$ relative to $\mathcal{M}_s$. Indeed, the difference between the CDFs under the two models is apparent at the 0.05 significance level, according to the two sample Kolmogorov-Smirnov test [*Massey*, 1951]. While it appears that there is only a slightly higher probability of large $\hat{\gamma}$ values under $\mathcal{M}_n$, extreme percentiles of the LPIII distribution are very sensitive to γ. Previous studies suggest γ should be with within ± 1.4 [*Reis and Stedinger*, 2005]. Therefore, estimation of $\boldsymbol{\theta}_n$ should not be based solely on MLE in the LPIII model. Incorporating regional information through an informative prior is an attractive methodology to restrict the inference of γ to realistic values.

Lastly, let us discuss how the posterior samples derived from the MCMC procedure can be used for model comparison and selection. Several common model comparison metrics include the small sample Akaike Information Criterion (AIC_c), the Bayesian Information Criterion (BIC), and the Deviance Information Criterion (DIC) [*Burnham and Anderson*, 2003; *Spiegelhalter et al*., 2002]. Under all three metrics, the goodness of fit term depends on $L(\boldsymbol{\theta}_j|\mathbf{X}, \mathcal{M}_j)$, whereas each metric applies a different penalty for model complexity. The model which minimizes the metric value is preferred given the data in the fitting procedure, and the difference in the metric values between competing models is used for model selection. The differences in AIC_c, BIC, and DIC values between $\mathcal{M}_s$ and $\mathcal{M}_n$ for the Smith River fitting period are 23.7, 22.5, and 22.1, respectively, each supporting $\mathcal{M}_n$. Rules of thumb for interpreting these differences are given by *Burnham and Anderson* [2003]. Based on the listed differences of each model selection metric, we conclude that there is

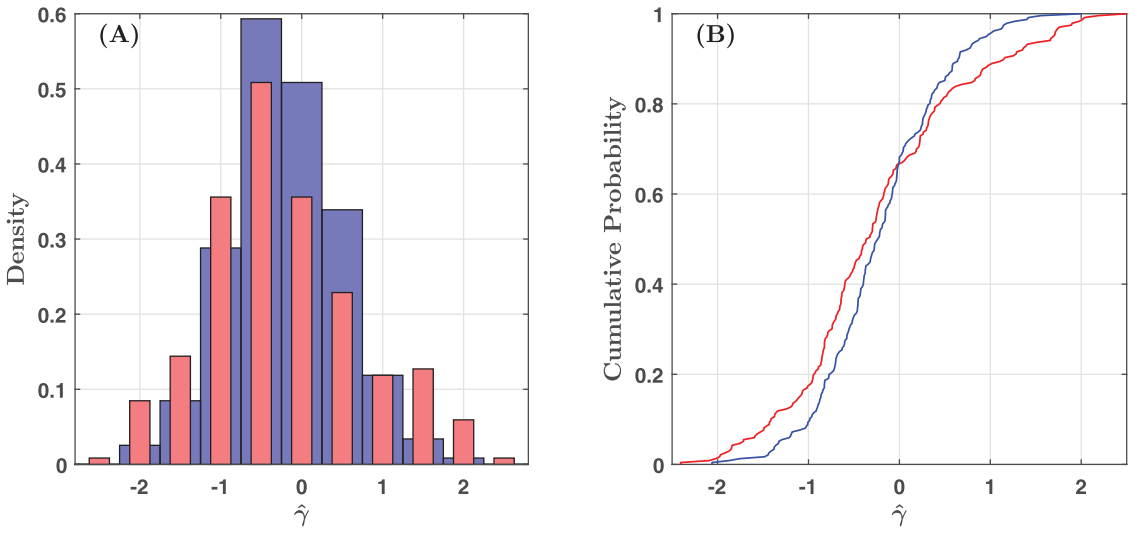

Figure 5. (a) Empirical density of $\hat{\gamma}$ under the NS model (*red* histogram) and the ST model (*blue* histogram) based on the estimates from fitting periods where $\hat{\alpha}$ was one standard deviation away from the mean of zero (236 total). The circumflex denotes the MLE of γ and α. (b) Empirical CDF of $\hat{\gamma}$ under the NS model (*red* line) and ST model (*blue* line). The CDF of $\hat{\gamma}$ under the NS model is different from the ST model counterpart, indicating that the likelihood function favors larger values of $\hat{\gamma}$ (i.e., ± 1.4) under the NS model. Indeed, there is a difference between the two CDFs at the 0.05 significance level according to the two sample Kolmogorov-Smirnov test [*Massey*, 1951].

essentially no empirical support for $\mathcal{M}_s$ relative to $\mathcal{M}_n$ given the data in the Smith River fitting period. Similar conclusions are drawn for the other data records.

Figure 6 shows the differences in fit metric values between $\mathcal{M}_s$ and $\mathcal{M}_n$ for fitting periods with detected trends (first row) and without detected trends (second row). Based on the differences presented in Figure 6, it is clear that the presence of a trend in the fitting period causes AIC_c, BIC, and DIC to significantly favor $\mathcal{M}_n$. Indeed, even for records without a significant trend, fit metrics can still indicate considerably less support for $\mathcal{M}_s$ relative to $\mathcal{M}_n$ because of the ST model's inferior ability to maximize the likelihood function. Model selection based solely on goodness of fit metrics may cause an analyst to sensibly select $\mathcal{M}_n$ for prediction given the presence of a trend in the data. However, goodness of fit metrics are primarily dependent on $L(\boldsymbol{\theta}_j|\mathbf{X}, \mathcal{M}_j)$, even if different penalties are applied for model complexity. Therefore, model selection rests on the ability to reproduce the historic record. Yet this begs the question of whether the ability to reproduce historic records is relevant for NS model selection. When comparing multiple ST models, this paradigm is useful, since we are not anticipating change in the historic distribution (by definition of stationarity). At least for engineering applications, the value of NS models should be dependent on their predictive ability, not their descriptive ability, as measured by goodness of fit. This issue of NS model selection and validation has been raised by other authors [*Renard et al*., 2013], and *Burnham and Anderson* [2003] note that their proposed guidelines for model selection cannot be expected to hold if observations are not independent. In truth, the predictive capabilities of a model can only be properly quantified using data that were not used in the fitting procedure. In this study, the out-of-sample predictions of the competing models are evaluated using Bayesian hypothesis testing.

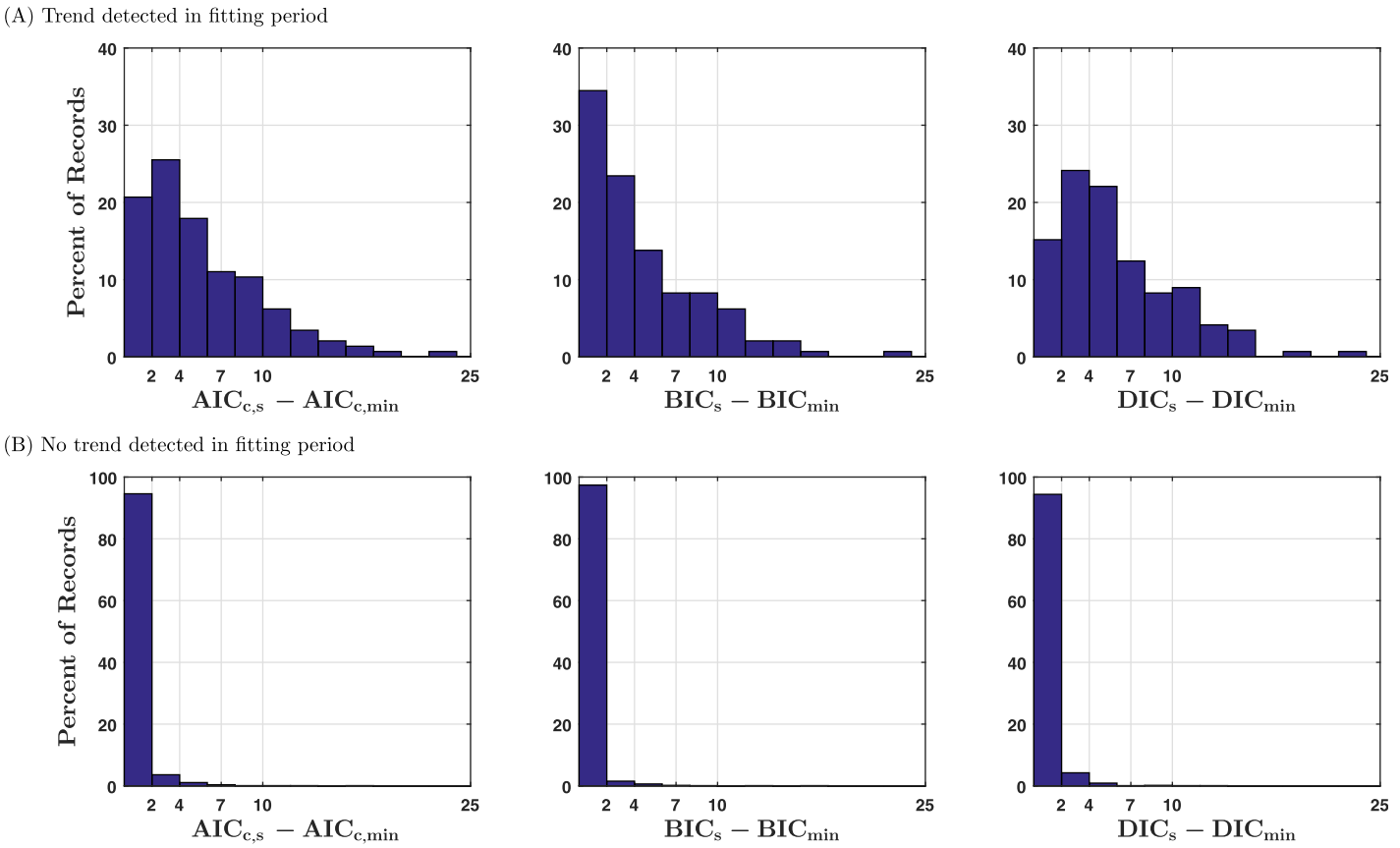

Figure 6. Comparison of model fit between the ST and NS models. (row a) Displays the differences in metric values for fitting periods with a detected trend (145 total) and (row b) displays the differences for fitting periods without detected trends (1105 total). Each column represents a different goodness of fit metric. The magnitude of the differences represents the level of empirical support for the ST model. Values from 0 to 2 in each figure indicate substantial support, values from 4 to 7 indicate considerably less, and values greater than 10 indicate essentially no empirical support the for ST model [*Burnham and Anderson*, 2003]. These results demonstrate that the presence of a trend substantially reduces the level of empirical support for the ST model relative to the NS model.

4.2. Bayesian Hypothesis Testing

Bayesian hypothesis testing depends on Bayes factor, which summarizes the evidence given by the data in favor of one statistical model relative to another [*Kass and Raftery*, 1995]. The "evidence" was introduced in equation (4), but here we are interested in determining the evidence given by the *out-of-sample* data, or $p(\mathbf{X}^*|\mathcal{M}_j)$. The out-of-sample data, $\mathbf{X}^*$, were previously defined in section 3. In this study, the evidence represents the probability of seeing the out-of-sample data that were actually observed under the competing models. Consequently, the evidence is extremely useful for model selection: models with larger values of $p(\mathbf{X}^*|\mathcal{M}_j)$ are statistically preferred for predicting the out-of-sample data. Model comparison is formally accomplished using the Bayes factor

$$B_{j,k} = 2\log_e\left[\frac{p(\mathbf{X}^*|\mathcal{M}_j)}{p(\mathbf{X}^*|\mathcal{M}_k)}\right] \tag{9}$$

where $B_{j,k}$ denotes the Bayes factor for $\mathcal{M}_j$ and against $\mathcal{M}_k$. Here we present a $\log_e$-scale formulation of the Bayes factor for simpler interpretation. Positive values of $B_{j,k}$ mean the evidence supports $\mathcal{M}_j$, and negative values support $\mathcal{M}_k$. The Bayes factor can be viewed as a predictive score, which measures the relative success of $\mathcal{M}_j$ and $\mathcal{M}_k$ at predicting $\mathbf{X}^*$. Interpretation of this Bayes factor formulation is given by *Kass and Raftery* [1995]. There are several important properties of Bayes factors and the evidence which make them especially useful for model selection.

First, the evidence automatically penalizes model complexity and parameter uncertainty. As the prior density diffuses through either larger parameter uncertainty or increasing dimensionality of the prior (i.e., increasing complexity), the value of the evidence will generally decrease (equation (4)). However, if increasing the dimensionality of the prior also causes the likelihood of the data to increase, the evidence could favor the more complex model. In this light, Bayes factor offers a formal means to measure if increased model complexity is justified by the data. Aside from favoring models with relatively concentrated priors, the evidence will also favor models where the prior parameter knowledge agrees with the data, which means that the likelihood function is maximized at or near the prior mode. This property allows different parameterizations of identical models to be evaluated according to Bayes factor. Overall, the Bayes factor is very sensitive to the choice of the prior, which is generally considered a downside of Bayesian hypothesis testing [*Kass and Raftery*, 1995]. However, sensitivity to the prior also creates opportunities for flexible applications.

4.2.1. Likelihood Function (Evaluation Period)

To compute Bayes factor, the evidence must be calculated under the competing models. This requires specification of the evaluation period prior and likelihood function for $\mathcal{M}_s$, $\mathcal{M}_u$, and $\mathcal{M}_n$, as well as evaluating the Bayes factor integral (equation (4)), which is a very challenging multi-dimensional integration. So let us first define the likelihood function, since it is relatively intuitive. To measure the predictive ability of the models, we must evaluate the likelihood function at the out-of-sample data

$$L(\boldsymbol{\theta}_j|\mathbf{X}^*, \mathcal{M}_j) = \prod_{i=1}^{n^*} p(X_i|\boldsymbol{\theta}_j, \mathcal{M}_j) \tag{10}$$

where n^* represents the number of observations in the evaluation data. Equation (10) is identical to the likelihood function of the fitting period, except here we are evaluating the PIII density function at the out-of-sample data. Also, $j = \{$"s," "n," "u"$\}$, since we are evaluating the predictions of all three models. When equation (10) is applied to calculate the likelihood of $\boldsymbol{\theta}_n$ under $\mathcal{M}_n$, again there is a time-dependent parameterization of the PIII density function (equation (7)), and t_i represents the time between the first observation of the fitting period and the ith observation of the evaluation period. We note this to emphasize that the trend parameter under $\mathcal{M}_n$ is extrapolated throughout the evaluation period.

Notice that the likelihood function distinguishes $\mathcal{M}_n$ from $\mathcal{M}_s$ and $\mathcal{M}_u$ but does not distinguish $\mathcal{M}_u$ from $\mathcal{M}_s$. Remember, under both $\mathcal{M}_s$ and $\mathcal{M}_u$, we predict future discharges are distributed according to a ST LPIII model. This justifies identical likelihood function evaluations. However, the parameters of the ST LPIII model are different under $\mathcal{M}_s$ and $\mathcal{M}_u$, since $\boldsymbol{\theta}_u$ was derived from the NS model. We can account for the difference between $\mathcal{M}_s$ and $\mathcal{M}_u$ through specification of the prior.

4.2.2. Prior Distributions (Evaluation Period)

In this study, our prior knowledge of the evaluation period model parameters is based on the posterior distributions inferred from the fitting period data. Even though observations were used to create this prior

knowledge, we can still consider this a prior belief about the evaluation period parameters, since the out-of-sample data were not used in the fitting procedure. This technique provides informative, joint priors describing the evaluation period model parameters, which allows for different parameterizations of the same model to be tested according to Bayes factor. Of course, we still need to formalize this prior knowledge as a probability distribution.

The evaluation period priors under $\mathcal{M}_s$, $\mathcal{M}_u$, and $\mathcal{M}_n$ were defined by a multivariate Gaussian mixture model (GMM) fit to the posterior samples of $\boldsymbol{\theta}_s$, $\boldsymbol{\theta}_u$, and $\boldsymbol{\theta}_n$, respectively. A GMM was chosen to describe the evaluation period priors since GMMs can represent a large class of distributions and approximate arbitrarily shaped densities. This property is ideal for approximating the joint densities of the posterior samples, since no assumption is required regarding the shape of marginal posteriors. In this study, the hyperparameters of the GMM were inferred through the Expectation Maximization (EM) algorithm. The number of Gaussians in the mixture, K, was determined by iteratively increasing the number of Gaussian components until $\mathrm{BIC}_K - \mathrm{BIC}_{K+1} < 2$. For more on mixture models and the EM algorithm, please see *Picard* [2007].

An example of the GMM fit to the posterior samples of the Smith River fitting period is shown in Figure 7. Figure 7 shows marginal prior densities of $\boldsymbol{\theta}_n$ (top row), $\boldsymbol{\theta}_s$ (middle row), and $\boldsymbol{\theta}_u$ (bottom row), although the fitted GMM also defines the joint density of parameter values. Black lines represent the prior density defined by the GMM, and the histograms are the posteriors from the Smith River fitting period. There are several important properties of the evaluation period priors that are highlighted by Figure 7. First, notice the difference in the prior modes under $\mathcal{M}_s$ compared to $\mathcal{M}_u$. Under $\mathcal{M}_s$, the prior favors lower values of μ relative to the μ_n prior density. This is because $\mathcal{M}_s$ assumes no change in the location of the distribution, so the out-of-sample μ should be similar to the stationary μ inferred from the fitting period. Conversely, under $\mathcal{M}_u$, a persistent shift in the distribution has occurred, so predictions are best made with the updated

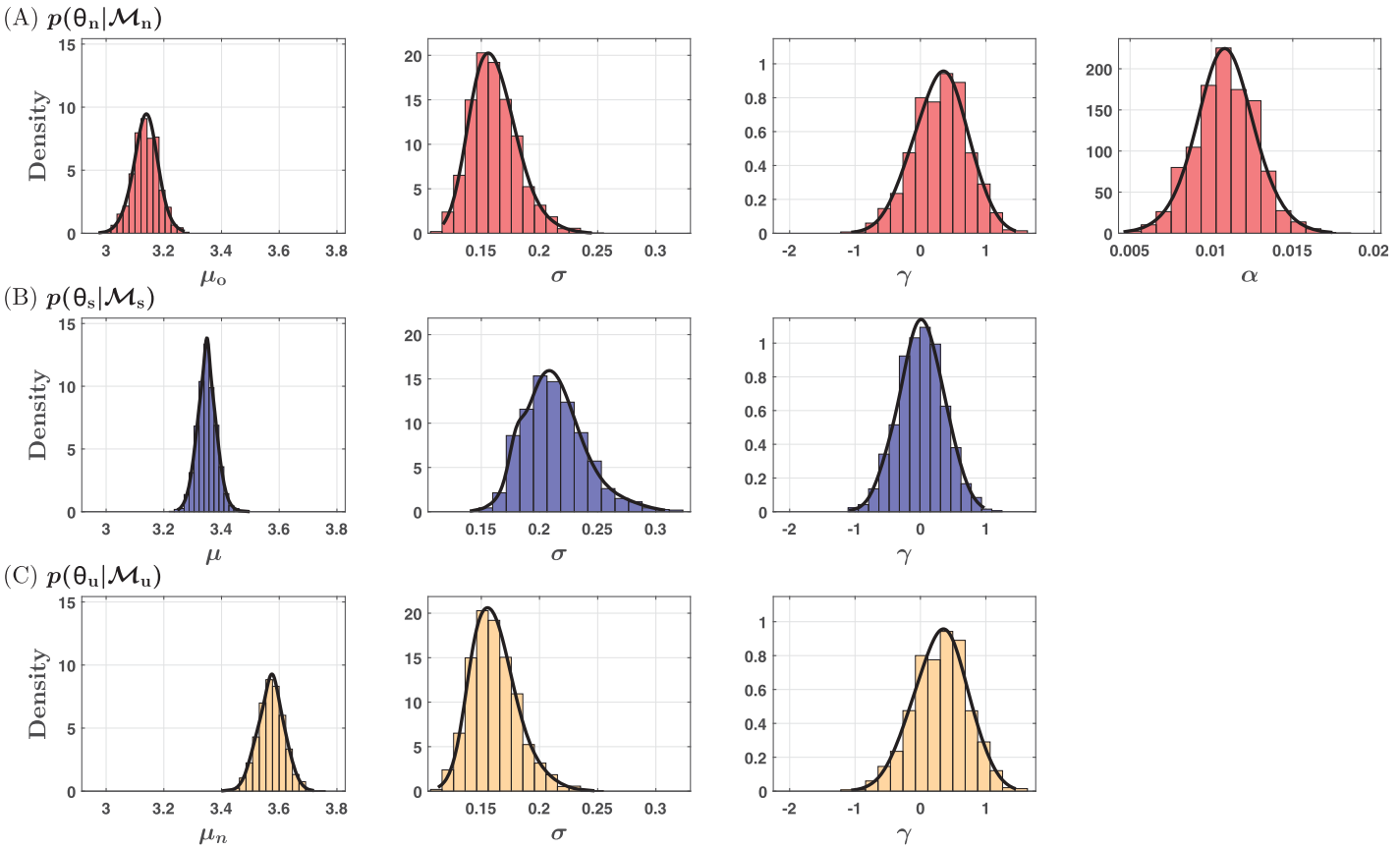

Figure 7. Prior parameter density for the evaluation period (*black* lines) under (row a) $\mathcal{M}_n$, (row b) $\mathcal{M}_s$, and (row c) $\mathcal{M}_u$ defined by the GMM fit to the posterior distributions of the Smith River fitting period (*red*, *blue*, and *gold* histograms). The additional complexity under $\mathcal{M}_n$ is accounted for through its four dimensional prior. Increased parameter uncertainty of μ_n is represented by its lower and more dispersed prior density relative to μ.

(larger) μ_n. The important concept here is that the predictions under $\mathcal{M}_s$ and $\mathcal{M}_u$ are set apart by their respective priors. In this study, the evidence will generally favor the model where the prior mode is closest to the out-of-sample μ and σ. However, this generalization does not consider how parameter uncertainty affects the model comparisons.

Notice also that the posterior samples of μ_n are significantly more dispersed than μ. Since samples of μ_n depend on the uncertain intercept and trend parameter of the linear model (equation (2)), they are understandably more dispersed than μ, which simply represents the mean of the fitting period. Figure 7 demonstrates the larger parameter uncertainty under $\mathcal{M}_u$, as shown by the lower and more dispersed prior probability density of μ_n relative to μ. Therefore, according to Bayes factor, $\mathcal{M}_u$ will only be favored relative to $\mathcal{M}_s$ if the increased parameter uncertainty is offset by a substantially more accurate prediction of the out-of-sample μ and σ. Following similar reasoning, $\mathcal{M}_n$ will only be favored relative to $\mathcal{M}_s$ and $\mathcal{M}_u$ if two conditions are met. First, $\mathcal{M}_n$ must substantially improve the likelihood of $\mathbf{X}^*$ in order to offset the diffusion of prior density caused by the addition of the trend parameter, α. Second, the trend inferred from fitting period must be similar to the out-of-sample trend, or else $p(\boldsymbol{\theta}_n|\mathcal{M}_n)$ will not agree with $L(\boldsymbol{\theta}_n|\mathbf{X}^*, \mathcal{M}_n)$, thus significantly diminishing the evidence. Therefore, Bayes factor will evaluate if predictions under the NS model improve the statistical representation of the out-of-sample data enough to justify the additional complexity and uncertainty introduced by extrapolation of the NS trend parameter. The last step toward implementing the proposed test requires integration of equation (4).

4.2.3. Integration of the Bayes Factor Integral

Analytic evaluation of the Bayes factor integral is only possible for a narrow class of models. In the absence of analytic solutions, the evidence is approximated using a variety of methods which are reviewed by *Kass and Raftery* [1995]. In this study, Gaussian Mixture Importance Sampling (GAME) is applied for approximation of the evidence [*Volpi et al.*, 2017]. Essentially, GAME is a Monte Carlo integration technique where the efficiency is improved through importance sampling, and the importance distribution is defined by a GMM fit to a posterior sample. The GAME estimator was calculated for each record and model by (1) generating another posterior sample with DREAM$_{(ZS)}$ using the evaluation period likelihood function and priors defined in section 4.2, (2) fitting a GMM to the second generated sample using the EM algorithm, and (3) applying the standard importance sampling integration strategy presented by *Kass and Raftery* [1995] using the fitted GMM as the importance distribution. The strength of the GAME estimator is that it is free of theoretical assumptions regarding the shape of the posterior, and it is computationally tractable through use of the importance distribution.

5. Results: Evaluation of Predictive Ability

Let us begin by revealing the evaluation period of the Smith River record. Figure 8 shows the data in the full Smith River record (Figure 8a), as well as predictions derived from the maximum a posteriori (MAP) parameter estimates compared to the out-of-sample density (Figures 8b–8d). Undeniably, the trend that was detected in the fitting period did not persist in the evaluation period. Thus, predictions made under both $\mathcal{M}_n$ and $\mathcal{M}_u$ drastically overestimate the out-of-sample mean. Bayes factor reflects these poor predictions by very strongly supporting $\mathcal{M}_s$ relative to both $\mathcal{M}_n$ and $\mathcal{M}_u$ (values given in Figure 8c, interpretation based on *Kass and Raftery* [1995]). Under $\mathcal{M}_s$, the trend that occurred throughout the fitting period was adequately represented through a large variance parameter, reflecting the variability of the hydrologic system. The apparent increase in discharge magnitude from 1932 to 1972 could be considered long-term persistence and not a significant "trend," or the increase could be considered a significant trend which simply did not persist throughout the prediction period. A proper physical understanding of the relationship between climatology, land use, and flood discharges in the Smith River watershed is needed to elaborate further. Indeed, the detection of a trend alone offers little insight toward persistence, and the significance of detected trends is particularly uncertain for hydroclimatic data because dependency structures are poorly understood [*Cohn and Lins*, 2005; *Lins and Cohn*, 2011]. This point is supported not only by previous studies and Smith River watershed but also by the results of our evaluation for other records tested.

Figure 9 summarizes the Bayes factor results for all records tested (1250 total). The pie charts display the Bayes factor value as a percentage of records tested. The columns compare the evidence between different models, or Bayes factor for $\mathcal{M}_n$ and against $\mathcal{M}_s$ (Figure 9a), for $\mathcal{M}_u$ and against $\mathcal{M}_s$ (Figure 9b), and for $\mathcal{M}_n$ and against $\mathcal{M}_u$ (Figure 9c). For example, blue to green colors in Figures 9a and 9b represent records

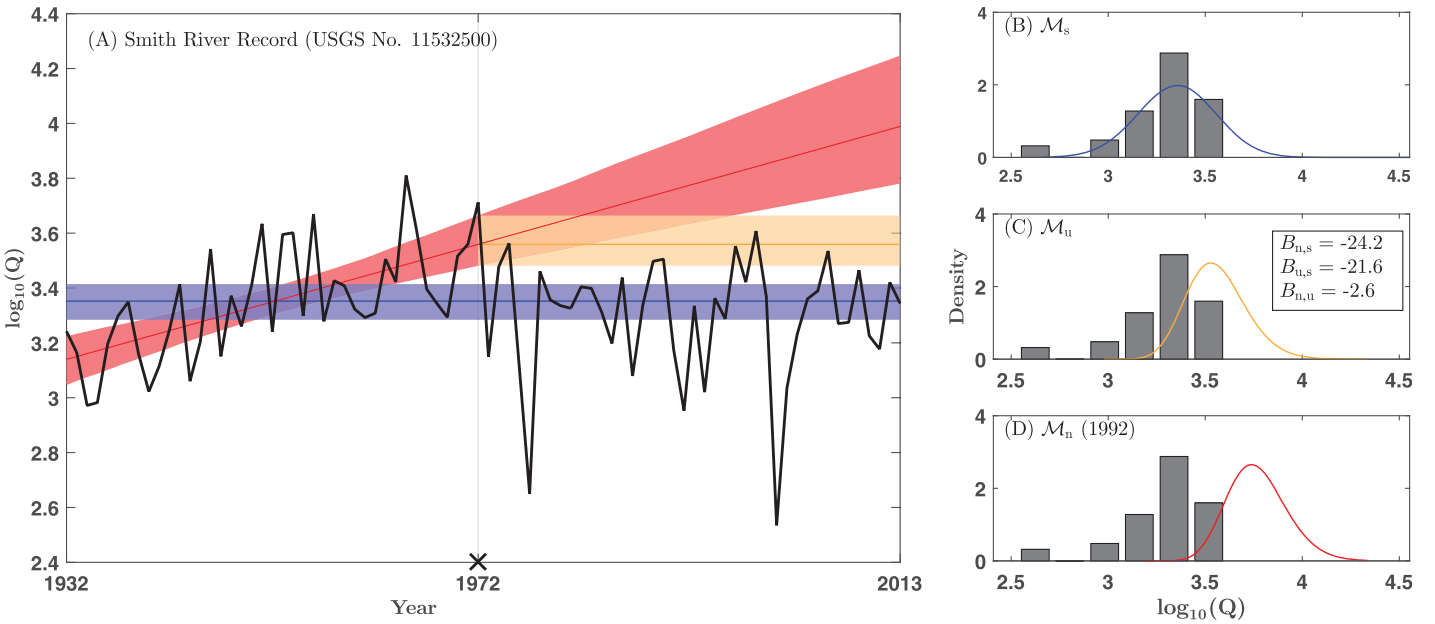

Figure 8. (a) The MAP estimate of the LPIII mean under $\mathcal{M}_s$ (*blue* line), $\mathcal{M}_u$ (*gold* line), and $\mathcal{M}_n$ (*red* line) shown over the full record length (*black* line). The colored shading represents the respective 95% credible intervals of the the LPIII mean. (b–d) Predictions of out-of-sample density under $\mathcal{M}_s$ (*blue* line), $\mathcal{M}_u$ (*gold* line), and $\mathcal{M}_n$ (*red* line) derived from the MAP parameter estimate. The *black* histograms show the empirical density in the evaluation period. For the Smith River record, $\mathcal{M}_s$ most accurately predicted the out-of-sample data, which is reflected by $B_{n,s}$ and $B_{u,s}$ shown in Figure 8c.

where the evidence favors $\mathcal{M}_s$ for prediction, whereas yellow to red colors in Figures 9a and 9c favor $\mathcal{M}_n$ for prediction. Interpretation of Bayes factor shown in Figure 9 legend is based on *Kass and Raftery* [1995]. Several important results are highlighted by examining the Bayes factor results over all records tested. First,

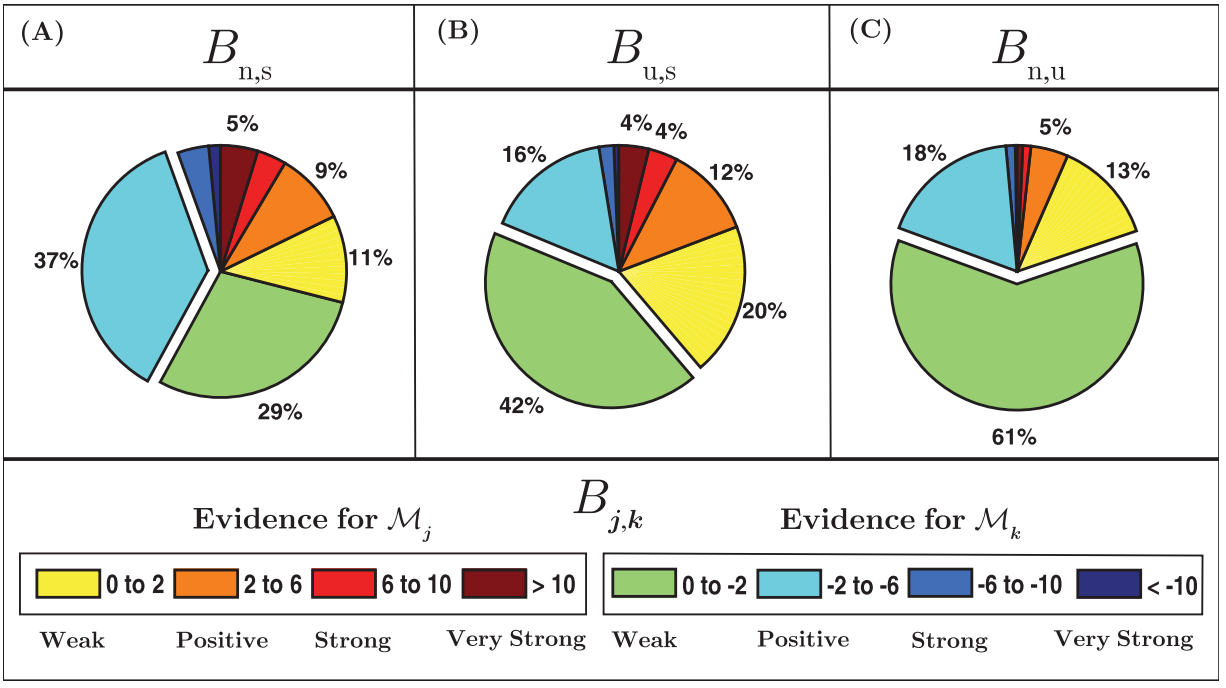

Figure 9. Bayes factor value as a percentage of all records tested (1250 total). (a) Bayes factor for $\mathcal{M}_n$ and against $\mathcal{M}_s$. (b) Bayes factor for $\mathcal{M}_u$ and against $\mathcal{M}_s$. (c) Bayes factor for $\mathcal{M}_n$ and against $\mathcal{M}_u$. *Yellow* to *red* values support $\mathcal{M}_j$, which is $\mathcal{M}_n$ in Figures 9a and 9c, and $\mathcal{M}_u$ in Figure 9b. *Green* to *blue* colors support $\mathcal{M}_k$, which is $\mathcal{M}_s$ in Figures 9a and 9b and $\mathcal{M}_u$ in Figure 9c. Results in Figure 9a show that the evidence supports predictions under $\mathcal{M}_s$ relative to $\mathcal{M}_n$ for about 70% of records tested, with 40% of records exhibiting at least positive evidence for $\mathcal{M}_s$. Results in Figure 9b show that predictions under $\mathcal{M}_s$ are also preferred relative to $\mathcal{M}_u$, except here the majority of evidence only weakly supports $\mathcal{M}_s$. Interpretation of the Bayes factor values is based on *Kass and Raftery* [1995].

Figure 9a shows that evidence favors $\mathcal{M}_s$ relative to $\mathcal{M}_n$ for roughly 70% of all records tested, with 40% of records exhibiting at least positive evidence toward $\mathcal{M}_s$. This means that the ST model was a better predictor of the out-of-sample data relative to the NS model extrapolations for about 70% of records tested. Bayes factor between $\mathcal{M}_s$ and $\mathcal{M}_u$ yield similar results (Figure 9b), except here the majority of tests offer only weak evidence toward $\mathcal{M}_s$ for prediction. Second, notice also that the evidence is not strongly in favor of any competing model when we examine the record pool as a whole, with the majority of the evidence falling within the weak to positive range. This is not surprising, since 88% of the records tested do not exhibit a statistically significant trend in the fitting period. Therefore, predictions under both $\mathcal{M}_n$ and $\mathcal{M}_u$ will be similar to $\mathcal{M}_s$ for most records. When models produce similar representations of data, Bayes factor naturally favors the simplest, so it is not surprising that the ST model is the apparent favorite for prediction over all records tested. However, this result does raise an important point: $\mathcal{M}_s$, or predictions of flood discharges using a ST model, should remain the default technique. The majority of flood records in the United States do not exhibit a statistically detectable trend, and therefore the trend parameter of the NS model is simply a nuisance parameter in most situations. Now let us examine records where NS and uST model predictions are substantially different from the ST model.

Figure 10 summarizes the results for subsets of the record pool with detected trends in the fitting period. The subsets shown in Figure 10 include records with a detected trend in the fitting period and DIC significantly favoring the NS model, or $\mathrm{DIC_s} - \mathrm{DIC_n} > 2$ (top row). The middle row shows the results for records with detected trends and discharges affected by land use changes, regulation, or diversion. The bottom row shows records where a trend was detected in the fitting period and full record length. Again, the columns compare the competing models. First, let us discuss records where trends were accompanied by fit metrics favoring the NS model (Figure 10a). Here the statistical metrics determined from the fitting period suggest that the ST assumption is violated and the NS model is a preferred alternative. Despite these metrics, Figure 10a shows that $\mathcal{M}_s$ is strongly preferred for prediction given records of this category. Bayes factor between $\mathcal{M}_s$ and $\mathcal{M}_n$ shows that the evidence is in favor of $\mathcal{M}_s$ for roughly 75% of these trending records. Moreover, about 40% exhibit at least *strong* evidence in favor of $\mathcal{M}_s$. Remarkably, comparison between Figures 9a and 10a illustrates that $\mathcal{M}_n$ is less preferred relative to $\mathcal{M}_s$ when a trend is detected and goodness of fit favors the NS model. Similarly, the evidence favors $\mathcal{M}_s$ relative to $\mathcal{M}_u$ for these records, although $\mathcal{M}_u$ did perform better than $\mathcal{M}_n$. This result confirms our earlier assertion that a statistically significant trend offers little insight toward persistence, and therefore little insight toward the predictive ability of the NS model. Furthermore, metrics of model fit measure the probability of reproducing the historic record, which is not related to trend persistence. To illustrate this important point further, Figure 11 shows the predictions under the three models compared to the data in the full record, given a trend and fit metrics favoring $\mathcal{M}_n$. Indeed, trends in the shown fitting periods are visually apparent and statistically significant, yet both $\mathcal{M}_n$ and $\mathcal{M}_u$ are poor predictors of the second half of the record. Thus, additional criteria are needed to confidently and reliably select NS models for prediction. Let us now examine records where trends were accompanied by indications of persistent change.

Figure 10b summarizes the results for trending fitting periods where anthropogenic influences were known to affect discharges. Records in Figure 10b were affected by upstream regulation, diversion, land use changes, or channelization during the fitting period. Again, we rely on USGS data flagging to find influenced records (section 3). Only 34 records tested meet this criteria, primarily because the first half of USGS records rarely contain data flags *and* significant trends. For these records, both $\mathcal{M}_u$ and $\mathcal{M}_n$ are preferred relative to $\mathcal{M}_s$. Notably, the evidence is *very strongly* in favor of $\mathcal{M}_u$ relative to $\mathcal{M}_s$ for more than a third of the records in Figure 10b, with 50% of records exhibiting at least *strong* evidence for $\mathcal{M}_u$. The relative success of the predictions derived from the NS model can be attributed to the permanent nature of anthropogenic watershed changes. Here we know the cause of the observed trend will persist without intervention, so it is much more likely for NS and uST predictions to be successful. Figure 12 shows successful predictions under $\mathcal{M}_u$ relative to $\mathcal{M}_s$ for selected records in Figure 10b. It is very important to note that $\mathcal{M}_n$ was preferred relative to $\mathcal{M}_u$ for only 2 of the 34 records in Figure 10b, and still the evidence only weakly supported $\mathcal{M}_n$. Thus, for records with a detected trend in the fitting period likely caused by anthropogenic influences, the uST model is preferred for prediction. This means that extrapolation of inferred NS model parameters is almost never preferred, even when the physical cause of the observed trend is known to be persistent. We emphasize this point through examination of one more subset of the record pool.

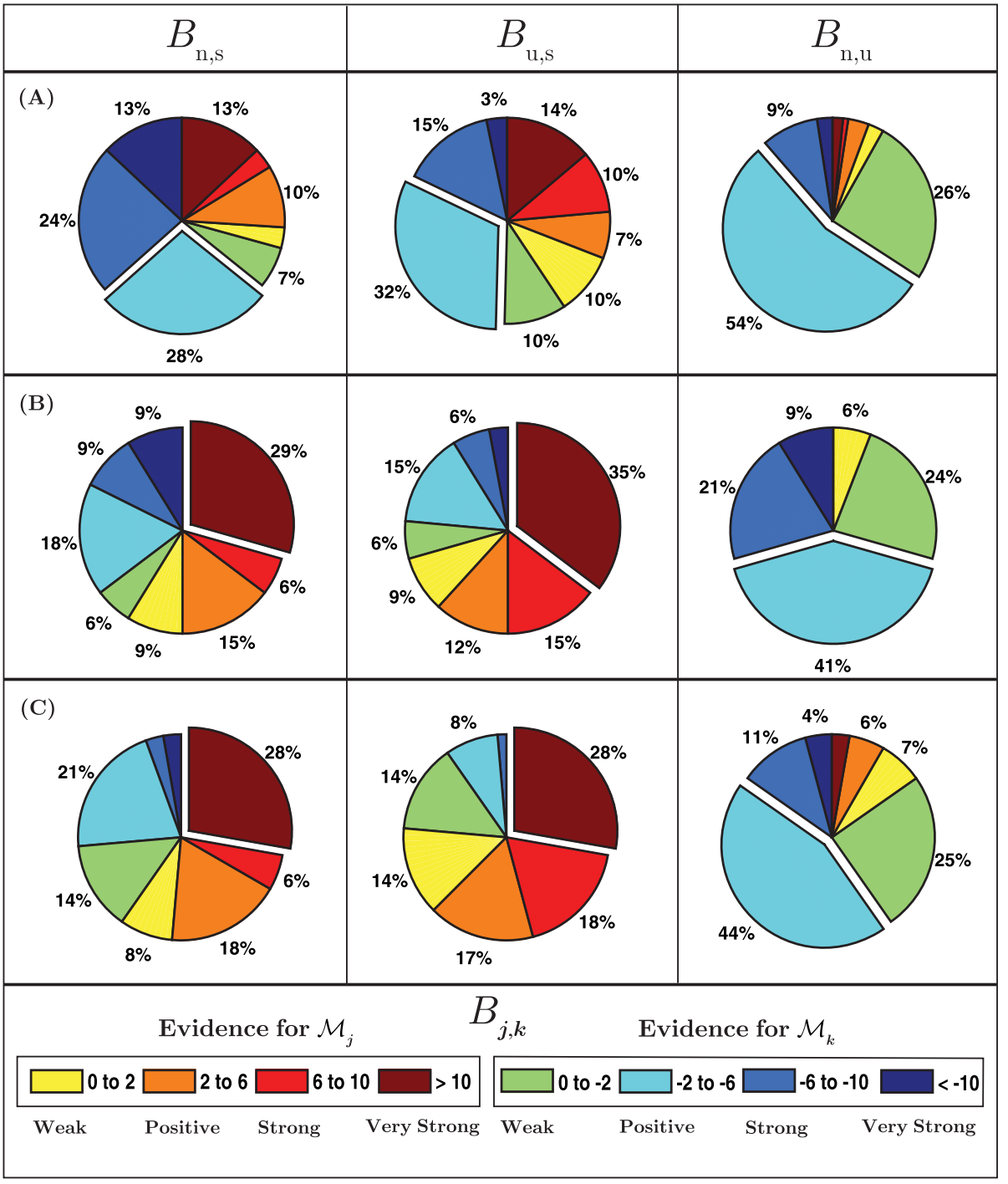

Figure 10. Bayes factor value as a percentage of records with detected trends in the fitting period and (a) DIC significantly favoring $\mathcal{M}_n$ (123 total), (b) discharges affected by land use changes, regulation, or diversion (34 total), and (c) trend also detected using full record length (72 total). The columns compare the evidence between competing models in the same manner as Figure 9. (a) Reveals that the presence of a trend and DIC favoring the NS model does not improve the relative success of $\mathcal{M}_n$ compared to all records tested in Figure 9. (b) Shows that $\mathcal{M}_u$ is strongly preferred for prediction relative to $\mathcal{M}_s$ and $\mathcal{M}_n$ if a trend is accompanied by physical watershed changes.

Figure 10c summarizes the results for records where the trend detected in the fitting period was also detected using the full record length. In other words, the monotonic trend detected in the fitting period persists in the evaluation period, although not necessarily at the same rate. In total, 72 records meet this criterion. Again, both $\mathcal{M}_u$ and $\mathcal{M}_n$ are preferred relative to $\mathcal{M}_s$. Figure 10c also shows that $\mathcal{M}_n$ only significantly improves predictions relative to $\mathcal{M}_u$ for about 10% of the records with an observed trend in the

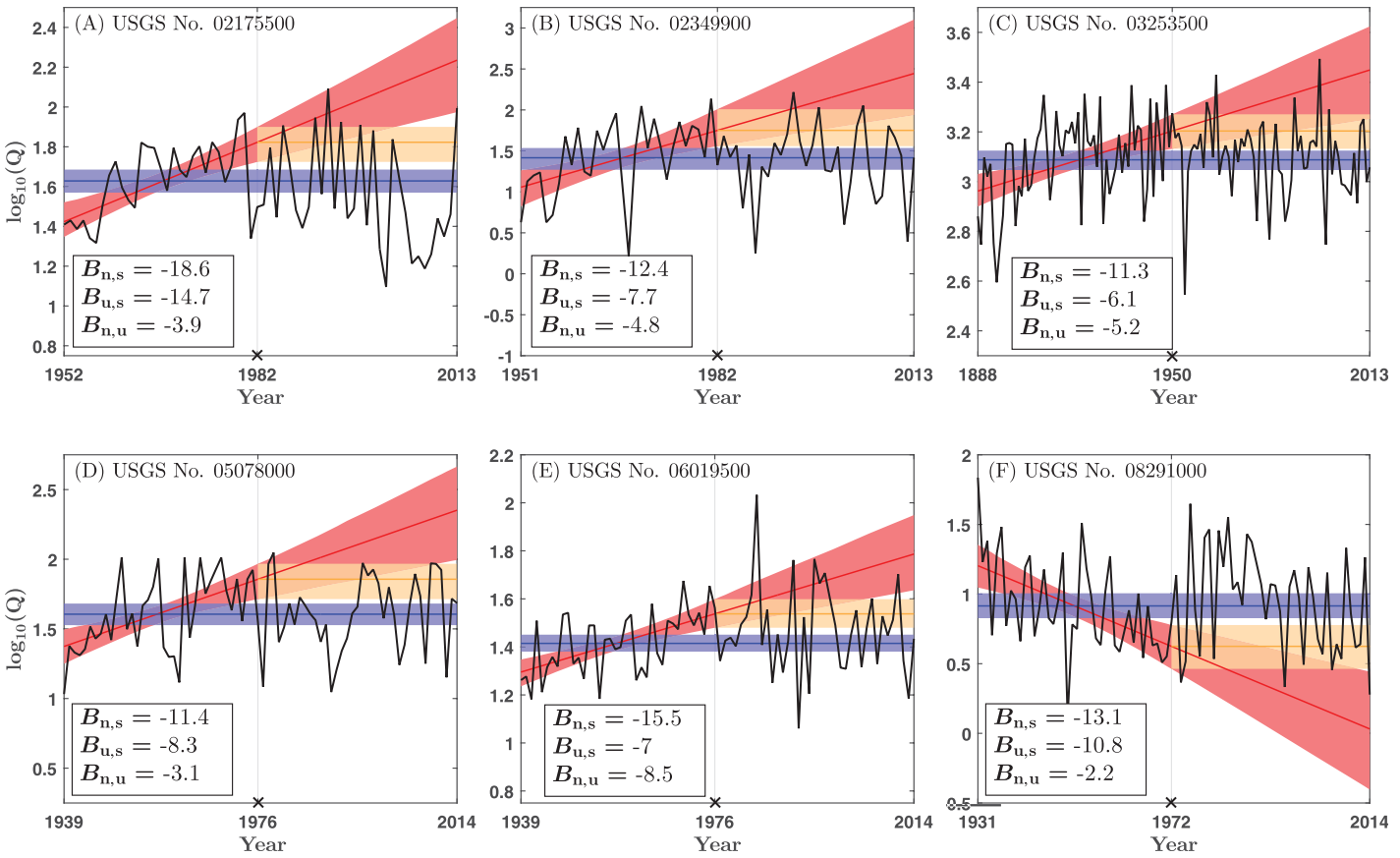

Figure 11. The MAP estimate of the LPIII mean under $\mathcal{M}_s$ (*blue* line), $\mathcal{M}_u$ (*gold* line), and $\mathcal{M}_n$ (*red* line) shown over the full data record (*black* line). The colored shading represents the respective 90% credible intervals of the LPIII mean, and the *black* cross denotes the end of the fitting period. The Bayes factor values are shown in the text boxes. Records (a–f) belong to category (a) of Figure 10, i.e., they exhibit a significant trend in the fitting period and DIC favors the NS model.

fitting period and full record length. The Bayes factor demonstrates that the NS model rarely enhances sufficiently the statistical description of the out-of-sample data (evaluation period) to justify its additional complexity and associated prediction uncertainty. This is most clearly illustrated in Figures 12e and 12f. The 90% prediction ranges of $\mathcal{M}_n$ (light red) cover much better the observed annual peak flows than their intervals in light orange from $\mathcal{M}_u$, which appear to be systematically biased. Nevertheless, the Bayes factor only weakly supports $\mathcal{M}_n$ relative to $\mathcal{M}_u$ in Figure 12e, and strongly supports $\mathcal{M}_u$ relative to $\mathcal{M}_n$ in Figure 12f. This finding may seem contradictory at first, but the mean of flood peak data in the evaluation period is within the 90% credible intervals of μ under $\mathcal{M}_u$. Therefore, an out-of-sample trend could be accounted for by making predictions with a conservative value of μ selected from within the uncertainty bounds under $\mathcal{M}_u$, and this approach is likely preferred relative to extrapolations of an estimated trend under $\mathcal{M}_n$. Appropriate quantiles of uST μ that are useful for prediction in the case of an out-of-sample (future) trend could be recommend via simulation in future studies. While often overlooked, the ST assumption appears to have some utility even in the context of NS [*Matalas*, 2012]. Before stating conclusions, let us discuss the implications and limitations of these results.

6. Discussion

Results from this study show that the extrapolation of parameter trends in the tested NS model does not significantly improve prediction, even when the observed trend persists into the future. However, results also show that extrapolation is not necessary to improve predictions relative to the traditional ST approach. The uST model, derived from the most recent in-sample NS model parameters, is preferred for prediction when a detected trend can be attributed to physical watershed changes (see *Viglione et al.* [2016] and

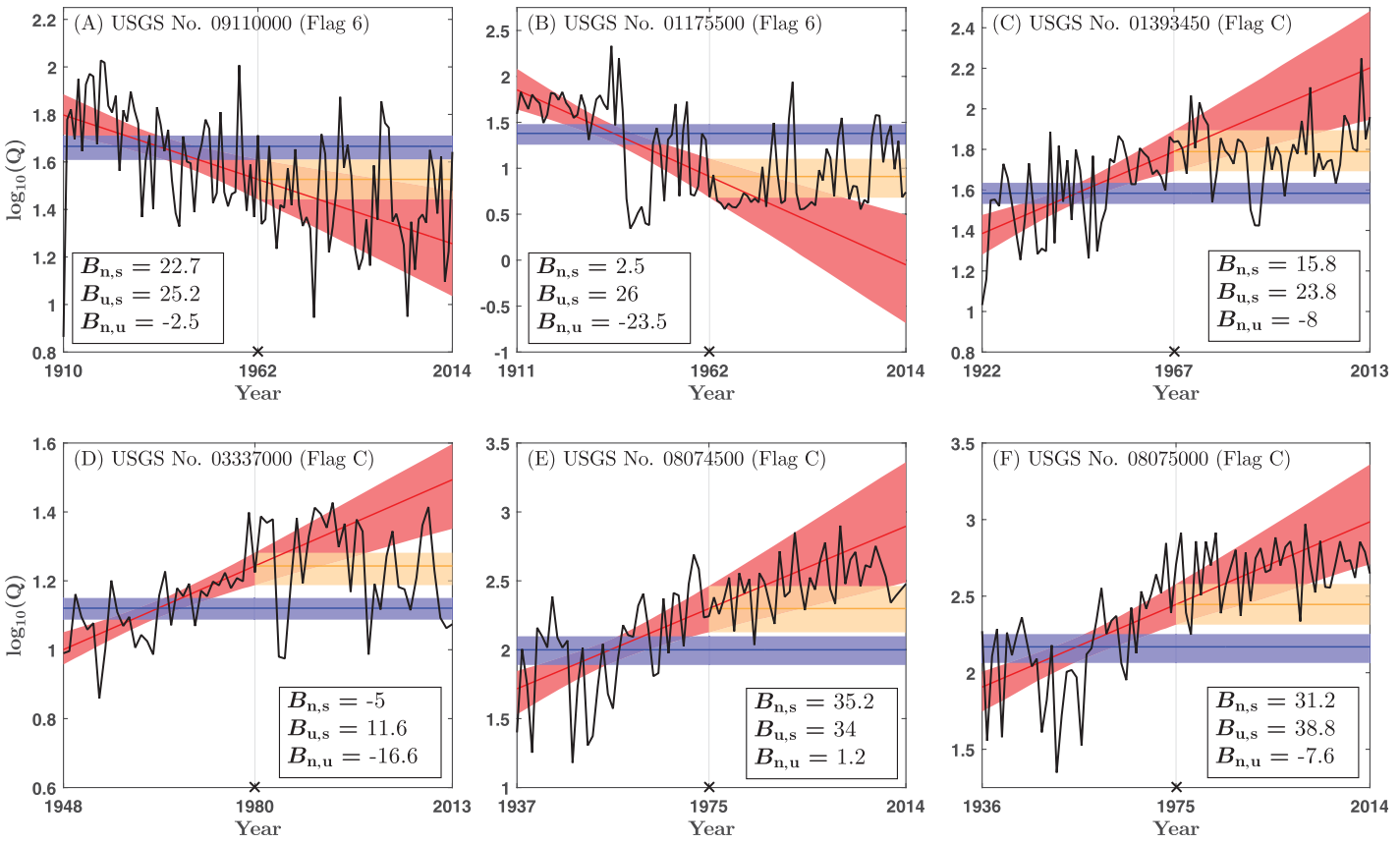

Figure 12. The MAP estimate of the LPIII mean under $\mathcal{M}_s$ (*blue* line), $\mathcal{M}_u$ (*gold* line), and $\mathcal{M}_n$ (*red* line) shown over the full data record (*black* line). The colored shading represents the respective 90% credible intervals of the LPIII mean, and the *black* cross denotes the end of the fitting period. The Bayes factor values are shown in the text boxes. Records (a–f) belong to category (b) of Figure 10, i.e., they exhibit a significant trend in the fitting period and discharges are affected by land use changes or channelization (flag C), upstream regulation, or diversion (flag 6).

references therein for recent NS attribution methods). Other authors have emphasized the importance of considering NS when the causes are well understood [*Lins and Cohn*, 2011], and *Vogel et al*. [2011] stressed that much greater attention should be given to anthropogenically influenced watersheds. Yet in the United States, there are no federal guidelines for FFA in watersheds where documented physical alterations have changed the local hydrology. Using NS models to update ST distributions is an attractive approach for the statistical treatment of these watersheds. Predictions under the uST distribution only assume that there has been a persistent change in the distribution of flood discharges, and our results show that even if the distribution continues to change, the uST distribution is likely preferred relative to extrapolations of the NS model parameters conditioned on time. Also, by making predictions within the ST paradigm, the concepts of return period and risk do not have to be extended to a fully NS framework for the design of hydraulic structures and hazard mapping. Until research shows *how* our prediction of flood peak distributions can be significantly improved by the extrapolation of NS model trend parameters, the ST assumption still holds for planning and design [*Montanari and Koutsoyiannis*, 2014].

For illustrative purposes, Figure 13b shows the return periods derived from the quantiles of the uST LPIII distribution using the full Whiteoak Bayou record near Houston, Texas for parameter inference (Figure 13a). The uncertainty in the predicted 100 year discharge (q_{99}) is also shown (Figure 13c), demonstrating the significant uncertainty in the estimate and the advantage of the Bayesian approach for parameter inference. We provide a MATLAB® program in the supporting information which can be used for inference of the uST distribution parameters from complete stream flow records, but we must emphasize several limitations. First, the Bayesian procedure applied herein measured the relative success of predicting the *overall* out-of-sample distribution, and not the precision of extreme quantiles. While the Bayesian approach is particularly useful for assessing the justified level of model complexity, traditional statistical techniques are more

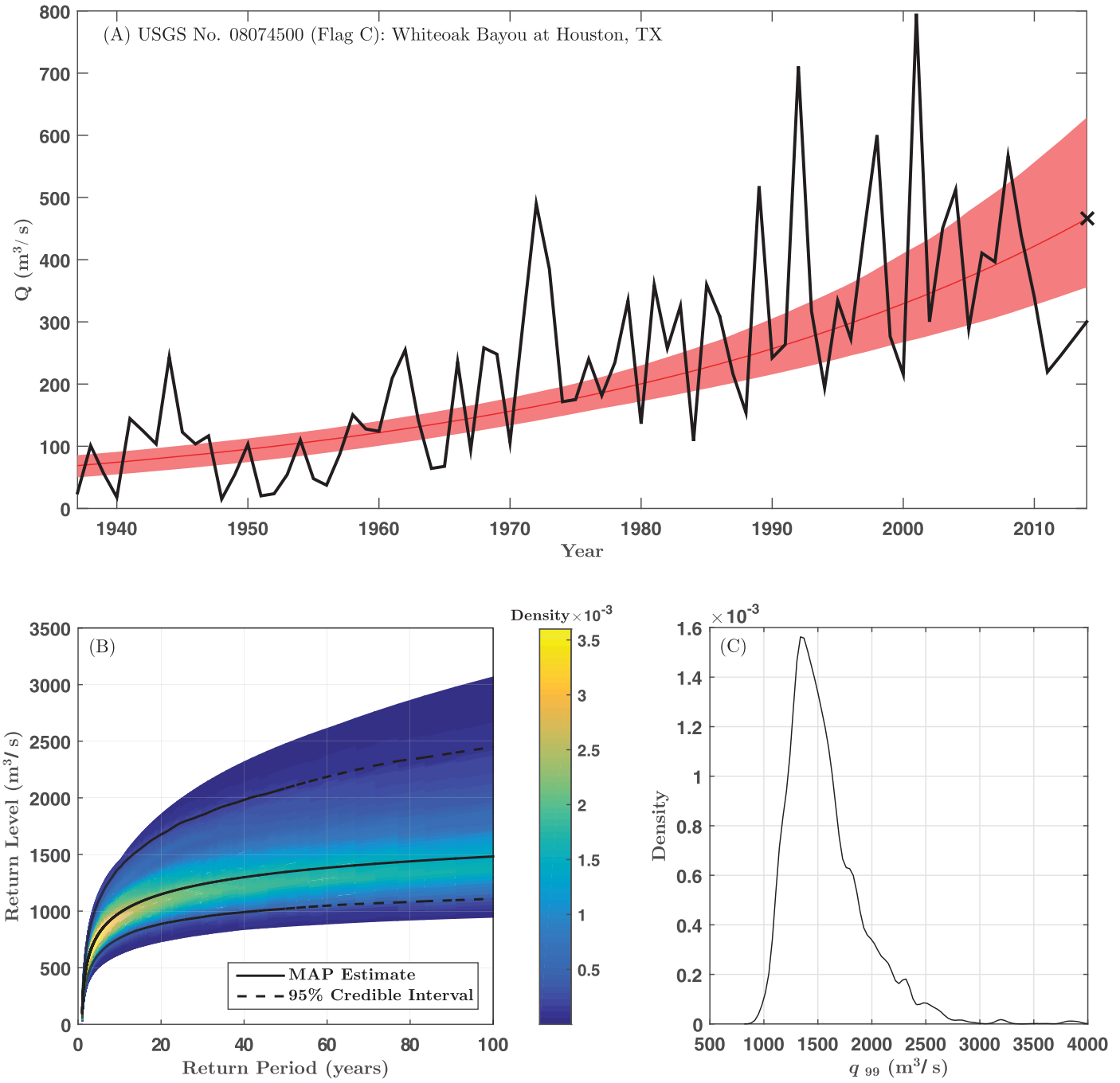

Figure 13. (a) The MAP estimate of the NS LPIII mean and 95% credible interval (*red* line and shading) inferred from the full record length (*black* line), shown outside of the log-space. The *black* cross denotes the mean of the uST distribution. (b) Return level versus return period plot derived from the uST distribution. The solid *black* line represents the MAP estimate of each return level, and the dashed *black* lines represent the 95% credible interval. The density of the return level estimates is shown on the color bar, which are readily available from the posterior sample of θ_u. (c) Distribution of the q_{99} estimates (or 100 year return level) under the uST model.

appropriate for testing the accuracy of specific quantiles. Under the traditional or frequentist approach, simulation is used to evaluate alternative quantile estimators in terms of bias and variance with respect to design flood estimates (see *Hosking and Wallis* [1997, section 6.4] for a detailed example). Therefore, the results show that the uST model is most appropriate among the tested models when NS behavior is obvious, but the uST model does not necessarily predict q_{99} accurately, for example. Future studies could

quantify the error in uST predictions of extreme percentiles through Monte Carlo simulation, similar to those conducted by *Yu et al*. [2015]. Second, although the uST model is *statistically* preferred for prediction relative to NS extrapolations, this result ignores the flood damages associated with under prediction. If we considered the losses associated with exceedances of design flood estimates, a more conservative prediction derived from the NS model extrapolations could be warranted, despite the additional complexity and uncertainty. Indeed, if an observed trend continues, it is unlikely for any derived uncertainty intervals of the uST quantiles to be accurate over a long time period (i.e., greater than 40 years). Therefore, we do not recommend basing design or insurance products on the return periods derived from the uST quantiles, but they can be used for evaluation of present flood risk in urbanizing or significantly altered catchments. While our results show that extrapolation of NS model parameters is not a statistically preferred alternative, this conclusion is somewhat limited to the tested NS model.

The primary limitation of this study is that only one NS model was tested for prediction. The tested NS LPIII model has notable advantages and disadvantages. From a practical point of view, modeling nonstationarity in the LPIII model is advantageous because of its widespread use in the United States and Australia for FFA. The LPIII distribution will continue to be utilized in the United States moving forward, since the forthcoming update to Hydrologic Bulletin 17B will not change the recommended distribution for FFA [*Lamontagne et al*., 2013]. Working with the LPIII distribution allows us to take advantage of the substantial prior knowledge of the distribution parameters. From a modeling point of view, the main advantage of the tested NS LPIII model is that changes in the mean and standard deviation of annual maximum discharges are captured with one additional trend parameter. This is a desirable property, since hydrologic nonstationarity would likely result in changes to both the location and scale [*Stedinger and Griffis*, 2011]. Therefore, the tested NS model is physically realistic and relatively simple. However, the linear trend in the log-mean equates to an exponential trend in the arithmetic mean. Therefore, extrapolation of the NS model parameters leads to nonlinear predictions of the changing mean. This is not desirable for out-of-sample predictions, which is reflected by the relative failure of NS extrapolations. It is possible that NS predictions would improve through extrapolation of a truly linear NS model. We also note that the sample size available for NS model parameter inference was particularly limited. Excluding half the available data was necessary for predictive analysis, but the annual maxima sampling method is wasteful of data. Sample size can be increased using a peak over threshold approach for extreme value sampling [*Renard et al*., 2006a; *Kyselý et al*., 2010], or taking advantage of regional trend information [*Cunderlik and Burn*, 2003; *Renard et al*., 2006b]. Both techniques would likely result in improved parameter inference and more reliable trend detection. Short-term trends and cycles that appear significant could also be avoided if the NS model incorporates additional information on historic floods [*Reis and Stedinger*, 2005; *Salinas et al*., 2016]. Another limitation of the tested model is that changes in the distribution of flood discharges are conditioned upon time.

By extrapolating a NS model for prediction using *time* as a covariate, we assume that the change in the distribution of flood discharges will continue at the same rate we have estimated from a limited sample. Furthermore, we assume a change in the distribution without a physical understanding of what caused the change. Extrapolations are therefore prone to failure since we have no insight toward the persistence of observed trends. Other authors have discussed the problems associated with NS trend extrapolation [*Koutsoyiannis and Montanari*, 2015], but the dominance of NS models in terms of traditional model selection metrics can be very misleading and actually cloud our judgment. The findings of this study largely support the claim by *Merz et al*. [2014] "Although statistical approaches have played and will play an important role, they have to be complemented by the search for the causal mechanisms and dominant processes in the atmosphere, catchment and river system that leave their fingerprints on flood characteristics." An understanding of the dominant mechanisms responsible for changes in flood discharge distributions allows for the development of physically based covariates [*Hall et al*., 2014; *Delgado et al*., 2014; *Steinschneider and Lall*, 2015; *Machado et al*., 2015; *Lima et al*., 2015]. Under this approach, the historic record is used to develop a relationship between the distribution and a physical process, such as atmospheric circulation patterns. Future changes to the distribution are then based on the inferred relationship and the prevalence of the physical process under general circulation model (GCM) based climate change projections. Here *the GCM is applied to provide insight toward the trend of important flood generating mechanisms, rather than to predict local and extreme precipitation events*. For example, *Šraj et al*. [2016] present a simple NS model where

changes in the frequency distribution are conditioned on annual precipitation. Design discharges can then be estimated for different projected values of annual precipitation that occur over the project lifespan. Indeed, physically based NS modeling is a promising direction toward meaningful extrapolation of NS model trend parameters, but continued research aimed at identifying important climate related signals in flood records is needed [*Archfield et al*., 2016]. We note that extrapolation of inferred trends should only be *considered* when there is strong evidence that the observed trend is likely to continue in the future. Examples include frequency analysis of temperature [*Cheng et al*., 2015], sea level extremes [*Mudersbach and Jensen*, 2010], or FFA in a watershed where urbanization is expected to continue in the future.

7. Conclusions and Recommendations

Results show that the ST model is preferred for out-of-sample prediction, overall. Even for records with a detected trend in the fitting period and goodness of fit metrics significantly favoring the NS model, the ST predictions are preferred relative to the uST and NS predictions. Therefore, the Mann-Kendall trend test applied to peak discharges and goodness of fit metrics alone are not sufficient to warrant the selection of a NS or uST model for prediction. This conclusion has been stressed by other authors [*Renard et al*., 2013; *Serinaldi and Kilsby*, 2015], and our results demonstrate empirical evidence toward this important assertion. For records with a detected trend and known physical watershed alterations (specifically USGS data flags 5/6 or C), the uST distribution is strongly preferred for prediction over the average evaluation period length of 40 years. Extrapolation of the NS model parameters is rarely preferred for prediction, even for records with a detected trend in the fitting period and full record. All things considered, we recommend the uST LPIII distribution for evaluation of current flood risk in watersheds with a detectable trend in annual maximum discharges that can be attributed to persistent, physical watershed changes. At this time, we do not recommend basing design or insurance products on predictions under the uST distribution because (1) the accuracy of extreme percentiles has not been comprehensively evaluated and (2) if the trend continues it is unlikely for any derived confidence bounds to encompass true quantiles over a long time period. We also recommend a fully Bayesian approach for parameter estimation so that (1) the bias of the NS skew estimate can be reduced with an informative prior and (2) parameter uncertainty is explicitly characterized. The supporting information provides a MATLAB® program based on the methods outlined in section 4.1 (applied to full record length) for inference of the ST, uST, and NS LPIII parameters. The program also includes basic postprocessing options. In conclusion, we recommend that future research should focus on the evaluation of predictions derived from promising NS models and the development of physically based covariates. Many stochastic models have been proposed, but we need to learn how to use them for prediction. Undoubtedly, practitioners are limited to standard statistical methods in the absence of accepted alternatives.

Appendix A: The Log-Pearson Type III Distribution

A record of block maxima discharges, $\{Q_1, \ldots, Q_n\}$, follows the LPIII distribution if random variable $\mathbf{X} = \log_{10}(\mathbf{Q})$ is distributed according to the PIII distribution. The PIII distribution parameters are defined by μ, σ, and γ, which represent the mean, standard deviation, and skewness of $\mathbf{X}$, respectively. The probability density function, $f(x)$, and cumulative distribution function, $F(x)$, are defined as

$$f(x) = \frac{(x-\xi)^{\epsilon-1} e^{-(x-\xi)/\beta}}{\beta^{\epsilon}\Gamma(\epsilon)} \tag{A1}$$

$$F(x) = G\left(\epsilon, \frac{x-\xi}{\beta}\right) \Big/ \Gamma(\epsilon) \tag{A2}$$

where $\epsilon = 4/\gamma^2$, $\beta = \frac{1}{2}\sigma\gamma$, and $\xi = \mu - 2\sigma/\gamma$ [*Hosking and Wallis*, 1997]. If $\gamma > 0$, then ξ is a lower bound $(\xi \leq x < \infty)$. If $\gamma < 0$, then ξ is an upper bound $(-\infty \leq x < \xi)$ and

$$f(x) = \frac{(\xi-x)^{\epsilon-1} e^{-(\xi-x)/\beta}}{\beta^{\epsilon}\Gamma(\epsilon)} \tag{A3}$$

$$F(x)=1-G\left(\epsilon,\frac{\xi-x}{\beta}\right)\Big/\Gamma(\epsilon) \tag{A4}$$

In the special case where $\gamma=0$, the distribution is normal and the range of x is $-\infty<x<\infty$. Here $\Gamma(\cdot)$ and $G(\cdot)$ represent the gamma and incomplete gamma functions, respectively. An analytical form of the inverse PIII distribution, $x(F)$, does not exist, thus it is often approximated to calculate discharges associated with specific quantiles, p, or return periods, T

$$\log_{10}(q_p)=\mu+\sigma K_p(\gamma) \tag{A5}$$

$$K_p(\gamma)=\frac{2}{\gamma}\left(1+\frac{\gamma n_p}{6}-\frac{\gamma^2}{36}\right)^3-\frac{2}{\gamma} \tag{A6}$$

where $\log_{10}(q_p)$ represents the pth quantile of the LPIII distribution, and n_p is the pth quantile of the standard normal distribution. $K_p(\gamma)$ is the frequency factor, or the pth quantile of the PIII distribution with mean 0, standard deviation of 1, and shape γ. The frequency factor is approximated by the Wilson and Hilferty transformation in equation (A6), which is accurate for $-2<\gamma<2$ and $0.01\le p\le 0.99$ or $1.01\le T\le 100$ [*Kirby*, 1972; *Reis and Stedinger*, 2005].

Because of the numerical difficulties in evaluating $\Gamma(\cdot)$ at large values of ϵ (γ values near 0), evaluation of the analytic PIII is often avoided. In this study, evaluation of the analytic PIII distribution near $\gamma=0$ is accomplished using Chebyshev approximations for $\log_e[\Gamma(\cdot)]$, which do not require evaluation of $\Gamma(\cdot)$ [*Cody and Hillstrom*, 1967]. Conveniently, $\log_e[\Gamma(\cdot)]$ can be computed using the MATLAB® function gammaln(), and evaluating the analytic PIII pdf is achieved in the log-space.

Acknowledgments
This work was made possible by the United States National Science Foundation (grant DMS-1331611), whose support we gratefully acknowledge. We also thank the anonymous reviewers for their thoughtful comments and suggestions. The data records used in this study can be found at URL: http://nwis.waterdata.usgs.gov/nwis/peak?

References

AghaKouchak, A., D. Easterling, K. Hsu, S. Schubert, and S. Sorooshian (Eds.) (2013), *Extremes in a Changing Climate*, 1st ed., pp. 39–92, Springer Sci. and Bus. Media, Dordrecht, Netherlands.

Apel, H., A. H. Thieken, B. Merz, and G. Blöschl (2006), A probabilistic modelling system for assessing flood risks, *Nat. Hazards*, *38*(1–2), 79–100.

Archfield, S., R. Hirsch, A. Viglione, and G. Blöschl (2016), Fragmented patterns of flood change across the United States, *Geophys. Res. Lett.*, *43*, 10,232–10,239, doi:10.1002/2016GL070590.

Begueria, S., and S. M. Vicente-Serrano (2006), Mapping the hazard of extreme rainfall by peaks over threshold extreme value analysis and spatial regression techniques, *J. Clim. Appl. Meteorol.*, *45*, 108–124, doi:10.1175/JAM2324.1.

Begueria, S., M. Angulo-Martinez, S. M. Vicente-Serrano, J. I. Lopez-Moreno, and A. El-Kenawy (2011), Assessing trends in extreme precipitation events intensity and magnitude using non-stationary peaks-over-threshold analysis: A case study in northeast Spain from 1930 to 2006, *Int. J. Climatol.*, *31*, 2102–2114, doi:10.1002/joc.2218.

Behrens, C. N., H. F. Lopes, and D. Gamerman (2004), Bayesian analysis of extreme events with threshold estimation, *Stat. Model. Int. J.*, *4*(3), 227–244.

Burnham, K. P., and D. R. Anderson (2003), *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, Springer Sci. and Bus. Media, New York.

Centre for Ecology and Hydrology (2008), *Flood Estimation Handbook*, Inst. of Hydrol., Wallingford, Oxfordshire, U. K.

Cheng, L., and A. AghaKouchak (2014), Nonstationary precipitation intensity-duration-frequency curves for infrastructure design in a changing climate, *Sci. Rep.*, *4*, 7093.

Cheng, L., A. AghaKouchak, and E. Gilleland (2014), Non-stationary extreme value analysis in a changing climate, *Clim. Change*, *127*, 353–369, doi:10.1007/s10584-014-1254-5.

Cheng, L., T. J. Phillips, and A. AghaKouchak (2015), Non-stationary return levels of CMIP5 multi-model temperature extremes, *Clim. Dyn.*, *44*(11–12), 2947–2963.

Cody, W. J., and K. E. Hillstrom (1967), Chebyshev approximations for the natural logarithm of the *gamma* function, *Math. Comp.*, *21*, 198–203, doi:10.1090/S0025-5718-67-99635-4.

Cohn, T. A., and H. F. Lins (2005), Nature's style: Naturally trendy, *Geophys. Res. Lett.*, *32*, L23402, doi:10.1029/2005GL024476.

Coles, S. (2001), *An Introduction to the Statistical Modeling of Extreme Values*, 1 ed., Springer, London, U. K.

Cooley, D. (2009), Extreme value analysis and the study of climate change, *Clim. Change*, *97*, 77, doi:10.1007/s10584-009-9627-x.

Cunderlik, J. M., and D. H. Burn (2003), Non-stationary pooled flood frequency analysis, *J. Hydrol.*, *276*(1), 210–223.

Delgado, J., B. Merz, and H. Apel (2014), Projecting flood hazard under climate change: An alternative approach to model chains, *Nat. Hazards Earth Syst. Sci.*, *14*(6), 1579–1589.

El Adlouni, S., T. Ouarda, X. Zhang, R. Roy, and B. Bobée (2007), Generalized maximum likelihood estimators for the nonstationary generalized extreme value model, *Water Resour. Res.*, *43*, W03410, doi:10.1029/2005WR004545.

European Commission (2007), Directive 2007/60/EC of the European Parliament and of the Council of 23 October 2007 on the assessment and management of flood risks. OJ L 288, 6.11.2007, pp. 27–34, Brussels.

Federal Emergency Management Agency (2009), Guidelines and specifications for flood hazard mapping partners, Appendix C: Guidance for riverine flooding analyses and mapping, *Tech. Rep. 13948*, Washington, D. C.

Gelman, A., and D. B. Rubin (1992), Inference from iterative simulation using multiple sequences, *Stat. Sci.*, *7*, 457–472.

Gilleland, E., and R. W. Katz (2011), New software to analyze how extremes change over time, *Eos Trans AGU*, *92*, 13, doi:10.1029/2011EO020001.

Griffis, V. W., and J. R. Stedinger (2007), Incorporating climate change and variability into bulletin 17B LP3 model, in *World Environmental and Water Resources Congress 2007*, Am. Soc. of Civ. Eng., Reston, Va., doi:10.1061/40927(243)69.
Hall, J., et al. (2014), Understanding flood regime changes in Europe: A state-of-the art assessment, *Hydrol. Earth Syst. Sci.*, *18*(7), 2735–2772.
Hallegatte, S., C. Green, R. J. Nicholls, and J. Corfee-Morlot (2013), Future flood losses in major coastal cities, *Nat. Clim. Change*, *3*, 802–806, doi:10.1038/nclimate1979.
Hirabayashi, Y., R. Mahendran, S. Koirala, L. Konoshima, D. Yamazaki, S. Watanabe, H. Kim, and S. Kanae (2013), Global flood risk under climate change, *Nat. Clim. Change*, *3*, 816–821, doi:10.1038/nclimate1911.
Hosking, J. R. M., and J. R. Wallis (1997), *Regional Frequency Analysis*, 1st ed., 200 pp., Cambridge Univ. Press, Cambridge, U. K.
Interagency Committee on Water Data (1982), Guidelines for determining flood flow frequency: Bulletin 17B, *Tech. Rep. 28*, Hydrol. Subcomm., Washington, D. C.
Jain, S., and U. Lall (2001), Floods in a changing climate: Does the past represent the future?, *Water Resour. Res.*, *37*(12), 3193–3205.
Jakob, J. (2013), Nonstationarity in extremes and engineering design, in *Extremes in a Changing Climate*, Springer, Dordrecht, Netherlands, doi:10.1007/978-94-007-4479-0 13.
Karl, T. R., J. M. Melillo, and T. C. Peterson (2009), *Global Climate Change Impacts in the United States*, Cambridge Univ. Press, New York.
Kass, R. E., and A. E. Raftery (1995), Bayes factors, *J. Am. Stat. Assoc.*, *90*(430), 773–795.
Katz, R. W. (2010), Statistics of extremes in climate change, *Clim. Change*, *100*, 71–76, doi:10.1007/s10584-010-9834-5.
Kendall, M. (1976), *Rank Correlation Methods*, 4th ed., Griffin, London.
Kirby, W. (1972), Computer-oriented Wilson-Hilferty transformation that preserves the first three moments and lower bound of the Pearson type 3 distribution, *Water Resour. Res.*, *8*(5), 1251–1254, doi:10.1029/WR008i005p01251.
Koutsoyiannis, D., and A. Montanari (2015), Negligent killing of scientific concepts: The stationarity case, *Hydrol. Sci. J.*, *60*, 1174–1183.
Kunkel, K. E., T. R. Karl, D. R. Easterling, K. Redmond, J. Young, X. Yin, and P. Hennon (2013), Probable maximum precipitation and climate change, *Geophys. Res. Lett.*, *40*, 1402–1408, doi:10.1002/grl.50334.
Kyselý, J., J. Picek, and R. Beranová (2010), Estimating extremes in climate change simulations using the peaks-over-threshold method with a non-stationary threshold, *Global Planet. Change*, *72*(1), 55–68.
Laloy, E., and J. A. Vrugt (2012), High-dimensional posterior exploration of hydrologic models using multiple-try $\text{DREAM}_{(zs)}$ and high-performance computing, *Water Resour. Res.*, *48*, W01526, doi:10.1029/2011WR010608.
Lamontagne, J. R., J. R. Stedinger, T. A. Cohn, and N. A. Barth (2013), Robust national flood frequency guidelines: What is an outlier?, paper presented at the Proceedings of the World Environmental and Water Resources Congress, Am. Soc. of Civ. Eng., Cincinnati, Ohio.
Lima, C. H., U. Lall, T. J. Troy, and N. Devineni (2015), A climate informed model for nonstationary flood risk prediction: Application to negro river at Manaus, Amazonia, *J. Hydrol.*, *522*, 594–602.
Lins, H. F., and T. A. Cohn (2011), Stationarity: Wanted dead or alive?, *J. Am. Water Resour. Assoc.*, *47*, 475–480.
Lopez, J., and F. Frances (2013), Non-stationary flood frequency analysis in continental Spanish rivers, using climate and reservoir indices as external covariates, *Hydrol. Earth Syst. Sci.*, *17*, 3189–3203, doi:10.5194/hess-17-3189-2013.
Machado, M. J., B. Botero, J. López, F. Francés, A. Díez-Herrero, and G. Benito (2015), Flood frequency analysis of historical flood data under stationary and non-stationary modelling, *Hydrol. Earth Syst. Sci.*, *19*, 2561–2576.
Madsen, H., D. Lawrence, M. Lang, M. Martinkova, and T. R. Kjeldsen (2013), A review of applied methods in Europe for flood-frequency analysis in a changing environment, COST Action ES0901: Flood frequency estimation methods and environmental change, technical report, Cent. for Ecol. and Hydrol., U. K., ISBN: 978-1-906698-36-2.
Mann, H. (1945), Nonparametric tests against trend, *Econometrica*, *13*, 245–259, doi:10.2307/1907187.
Martins, E. S., and J. R. Stedinger (2000), Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data, *Water Resour. Res.*, *36*(3), 737–744.
Massey, F. J., Jr. (1951), The Kolmogorov-Smirnov test for goodness of fit, *J. Am. Stat. Assoc.*, *46*(253), 68–78.
Matalas, N. C. (2012), Comment on the announced death of stationarity, *J. Water Resour. Plann. Manage.*, *138*(4), 311–312.
Merz, B., et al. (2014), Floods and climate: Emerging perspectives for flood risk assessment and management, *Nat. Hazards Earth Syst. Sci.*, *14*(7), 1921–1942.
Milly, P. C. D., R. T. Wetherald, K. Dunne, and T. L. Delworth (2002), Increasing risk of great floods in a changing climate, *Nature*, *415*, 514–517, doi:10.1038/415514a.
Min, S.-K., X. Zhang, F. W. Zwiers, and G. C. Hegerl (2011), Human contribution to more-intense precipitation extremes, *Nature*, *470*(7334), 378–381.
Montanari, A., and D. Koutsoyiannis (2014), Modeling and mitigating natural hazards: Stationarity is immortal!, *Water Resour. Res.*, *50*, 9748–9756, doi:10.1002/2014WR016092.
Mudersbach, C., and J. Jensen (2010), Nonstationary extreme value analysis of annual maximum water levels for designing coastal structures on the German North Sea coastline, *J. Flood Risk Manage.*, *3*(1), 52–62.
Nozdryn-Plotnicki, M., and W. Watt (1979), Assessment of fitting techniques for the log Pearson type 3 distribution using Monte Carlo simulation, *Water Resour. Res.*, *15*(3), 714–718.
Olsen, J. R. (2002), Climate change and floodplain management in the United States, *Clim. Change*, *76*, 407–426, doi:10.1007/s10584-005-9020-3.
Ouarda, T., and S. El-Adlouni (2011), Bayesian nonstationary frequency analysis of hydrological variables, *J. Am. Water Resour. Assoc.*, *47*, 496–505.
Peel, M. C., and G. Blöschl (2011), Hydrological modelling in a changing world, *Prog. Phys. Geogr.*, *35*(2), 249–261.
Perreault, L., J. Bernierand, B. Bobée, and E. Parent (2000), Bayesian change-point analysis in hydrometeorological time series. Part 2. Comparison of change-point models and forecasting, *J. Hydrol.*, *235*(3), 242–263.
Picard, F. (2007), An introduction to mixture models, *Res. Rep.* 7, *Stat. for Syst. Biol.*, Paris.
Pilgrim, D. (2001), *Australian Rainfall and Runoff: A Guide to Flood Estimation*, vol. *1*, Inst. of Eng., Barton, ACT, Australia.
Prosdocimi, I., T. R. Kjeldsen, and C. Svensson (2014), Non-stationarity in annual and season series of peak flow and precipitation in the UK, *Nat. Hazards Earth Syst. Sci.*, *14*, 1125–1144, doi:10.5194/nhess-14-1125-2014.
Raff, D. A., T. Pruitt, and L. D. Brekke (2009), A framework for assessing flood frequency based on climate projection information, *Hydrol. Earth Syst. Sci.*, *13*, 2119–2136, doi:10.1007/s10584-005-9020-3.
Read, L. K., and R. M. Vogel (2016), Hazard function analysis for flood planning under nonstationarity, *Water Resour. Res.*, *52*, 4116–4131, doi:10.1002/2015WR018370.

Reis, D. S., Jr., and J. R. Stedinger (2005), Bayesian MCMC flood frequency analysis with historical information, *J. Hydrol.*, *313*, 97–116, doi:10.1016/j.jhydrol.2005.02.028.
Renard, B., M. Lang, and P. Bois (2006a), Statistical analysis of extreme events in a non-stationary context via a Bayesian framework: Case study with peak-over-threshold data, *Stochastic Environ. Res. Risk Assess.*, *21*(2), 97–112.
Renard, B., V. Garreta, and M. Lang (2006b), An application of Bayesian analysis and Markov Chain Monte Carlo methods to the estimation of a regional trend in annual maxima, *Water Resour. Res.*, *42*, W12422, doi:10.1029/2005WR004591.
Renard, B., X. Sun, and M. Lang (2013), Bayesian methods for non-stationary extreme value analysis, in *Extremes in a Changing Climate*, pp. 39–95, Springer, Dordrecht, Netherlands.
Salas, J. D., and J. Obeysekera (2014), Revisiting the concepts of return period and risk for nonstationary hydrologic extreme events, *J. Hydrol. Eng.*, *19*, 554–568, doi:10.1061/(ASCE)HE.1943-5584.0000820.
Salas, J. D., B. Rajagopalan, L. Saito, and C. Brown (2012), Climate nonstationarity and water resources management, *J. Water Resour. Plann. Manage.*, *138*(5), 385–388.
Salinas, J. L., A. Kiss, A. Viglione, R. Viertl, and G. Blöschl (2016), A fuzzy Bayesian approach to flood frequency estimation with imprecise historical information, *Water Resour. Res.*, *52*, 6730–6750, doi:10.1002/2016WR019177.
Schaake, J., S. Cong, and Q. Duan (2006), The US MOPEX data set, *IAHS Publ.*, *307*, 9.
Serinaldi, F., and C. G. Kilsby (2015), Stationarity is undead: Uncertainty dominates the distribution of extremes, *Adv. Water Resour.*, *77*, 17–36, doi:10.1016/j.advwatres.2014.12.013.
Silva, A. T., M. Naghettini, and M. M. Portela (2015), On some aspects of peaks-over-threshold modeling of floods under nonstationarity using climate covariates, *Stochastic Environ. Res. Risk Assess.*, *30*, 207–224, doi:10.1007/s00477-015-1072-y.
Slack, J., and J. Landwehr (1992), Hydro-climatic data network (HCDN): A U.S. Geological Survey streamflow data set for the United States for the study of climate variations, *U.S. Geol. Surv. Open File Rep.*, *92-129*, 1–193.
Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde (2002), Bayesian measures of model complexity and fit, *J. R. Stat. Soc., Ser. B*, *64*, 583–639, doi:10.1111/1467-9868.00353.
Šraj, M., A. Viglione, J. Parajka, and G. Blöschl (2016), The influence of non-stationarity in extreme hydrological events on flood frequency estimation, *J. Hydrol. Hydromech.*, *64*, 426–437.
Srikanthan, R., and T. McMahon (1981), Log Pearson III distribution effect of dependence, distribution parameters and sample size on peak annual flood estimates, *J. Hydrol.*, *52*(1–2), 149–159.
Stedinger, J. R., and V. W. Griffis (2011), Getting from here to where? Flood frequency analysis and climate, *J. Am. Water Resour. Assoc.*, *47*, 506–513, doi:10.1111/j.1752-1688.2011.00545.x.
Steinschneider, S., and U. Lall (2015), A hierarchical Bayesian regional model for nonstationary precipitation extremes in Northern California conditioned on tropical moisture exports, *Water Resour. Res.*, *51*, 1472–1492, doi:10.1002/2014WR016664.
Strupczewski, W., V. P. Singh, and W. Feluch (2001), Non-stationary approach to at-site flood frequency modelling I. Maximum likelihood estimation, *J. Hydrol.*, *248*, 123–142, doi:10.1016/S0022-1694(01)00397-3.
Tramblay, Y., L. Neppel, J. Carreau, and K. Najib (2014), Non-stationary frequency analysis of heavy rainfall events in southern France, *Hydrol. Sci. J.*, *58*, 280–294, doi:10.1080/02626667.2012.754988.
Trenberth, K. E. (2011), Changes in precipitation with climate change, *Clim. Res.*, *47*(1–2), 123–138.
Viglione, A., B. Merz, N. Viet Dung, J. Parajka, T. Nester, and G. Blöschl (2016), Attribution of regional flood changes based on scaling fingerprints, *Water Resour. Res.*, *52*, 5322–5340, doi:10.1002/2016WR019036.
Villarini, G., and J. A. Smith (2009), Flood peak distributions for the eastern United States, *Water Resour. Res.*, *46*, W06504, doi:10.1029/2009WR008395.
Villarini, G., J. A. Smith, F. Serinaldi, J. Bales, P. D. Bates, and W. F. Krajewski (2009), Flood frequency analysis for nonstationary annual peak records in an urban drainage basin, *Adv. Water Resour.*, *32*, 1255–1266, doi:10.1016/j.advwatres.2009.05.003.
Vogel, R. M., C. Yaindl, and M. Walter (2011), Nonstationarity: Flood magnification and recurrence reduction factors in the United States, *J. Am. Water Resour. Assoc.*, *47*, 464–474.
Volpi, E., G. Schoups, G. Firmani, and J. A. Vrugt (2017), Sworn testimony of the model evidence: Gaussian mixture importance (GAME) sampling, *Water Resour. Res.*, *53*, doi:10.1002/2016WR020167, in press.
Vrugt, J. A. (2016), Markov Chain Monte Carlo simulation using the dream software package: Theory, concepts, and Matlab implementation, *Environ. Modell. Software*, *75*, 273–316.
Vrugt, J. A., C. T. Braak, C. Diks, B. A. Robinson, J. M. Hyma, and D. Higdon (2009), Accelerating Markov Chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling, *Int. J. Nonlinear Sci. Num.*, *10*(3), 273–290.
Wallis, J., D. Lettenmaier, and E. F. Wood (1991), A daily hydroclimatological data set for the continental United States, *Water Resour. Res.*, *27*(7), 1657–1663, doi:10.1029/91WR00977.
Yu, X., T. A. Cohn, J. R. Stedinger, K. Karvazy, and V. Webster (2015), Flood frequency analysis in the context of climate change, in *World Environmental and Water Resources Congress*, pp. 2376–2385, Am. Soc. of Civ. Eng., Reston, Va.