AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE

10.1002/2016WR019129

On the deterministic and stochastic use of hydrologic models

William H. Farmer[1] and Richard M. Vogel[2]

[1]National Research Program, U.S. Geological Survey, Denver, Colorado, USA, [2]Department of Civil and Environmental Engineering, Tufts University, Medford, Massachusetts, USA

Abstract Environmental simulation models, such as precipitation-runoff watershed models, are increasingly used in a deterministic manner for environmental and water resources design, planning, and management. In operational hydrology, simulated responses are now routinely used to plan, design, and manage a very wide class of water resource systems. However, all such models are calibrated to existing data sets and retain some residual error. This residual, typically unknown in practice, is often ignored, implicitly trusting simulated responses as if they are deterministic quantities. In general, ignoring the residuals will result in simulated responses with distributional properties that do not mimic those of the observed responses. This discrepancy has major implications for the operational use of environmental simulation models as is shown here. Both a simple linear model and a distributed-parameter precipitation-runoff model are used to document the expected bias in the distributional properties of simulated responses when the residuals are ignored. The systematic reintroduction of residuals into simulated responses in a manner that produces stochastic output is shown to improve the distributional properties of the simulated responses. Every effort should be made to understand the distributional behavior of simulation residuals and to use environmental simulation models in a stochastic manner.

Key Points:

- Deterministic use of environmental simulation models is inappropriate for operational hydrology
- Deterministic use of simulation models introduces distributional bias into results
- Prudent management of environmental resources requires stochastic use of simulation models

Correspondence to:
W. H. Farmer,
wfarmer@usgs.gov

Citation:
Farmer, W. H., and R. M. Vogel (2016), On the deterministic and stochastic use of hydrologic models, *Water Resour. Res.*, *52*, doi:10.1002/2016WR019129.

Received 26 APR 2016
Accepted 23 JUN 2016
Accepted article online 18 JUN 2016

1. Introduction

With increasing frequency, hydrologic models, and, more generally, environmental simulation models (ESMs), are being used for a myriad of applications in operational hydrology. Such uses include the design, planning, and management of water resource systems under changing climates, land use, and other anthropogenic shifts. Such models are often used in a deterministic fashion that ignores the model uncertainty associated with simulated responses. The primary goal of this work is to demonstrate how this approach leads to simulated responses that cannot reproduce the distribution of the observed responses. The impact of ignoring model uncertainty is shown to be magnified for hydrologic extremes and design-relevant products, such as design floods and droughts; problems that are likely to be exacerbated by climate change.

There is increasing and widespread attention given to uncertainty analysis of environmental and water resource simulation models as evidenced by the recent reviews by *Liu and Gupta* [2007], *Moradkhani and Sorooshian* [2008], *Matott et al*. [2009], *Montanari* [2011], *Beven* [2014], and *Mirzaei et al*. [2015]. However, all such literature focuses exclusively on estimation of uncertainty intervals for model responses. Although integration of the stochastic structure of ESM residuals is paramount to developing such uncertainty intervals, construction of uncertainty intervals does not remove the bias in design-relevant products such as design floods, droughts, and storage-yield curves. A goal of this study is to emphasize the importance and advantage of exploiting recent advances in the uncertainty analysis of ESMs for the purpose of improving the operational use of such models in water resources design, planning, and management. However, this discussion addresses only simulation error, leaving observational error, which can be substantial, aside.

Deterministic use and stochastic use refer to the way in which model output is used in subsequent applications. Here, the deterministic use of ESMs refers to the use of simulated responses as single, certain estimates of model response. Deterministic use employs simulated responses as a single realization of a simulated process and uses this single realization to derive any required design, planning, or management product. Stochastic use is an *ex post facto* solution, effectively a post-processing procedure, where model uncertainty is added to the model output by some means after the simulated response has been produced. The stochastic use of a statistical or deterministic model requires a Monte-Carlo process by which equally likely model output traces are produced. Note that, as in *Vogel* [1999], both statistical and deterministic models are viewed as equivalent in the sense that both types of models consist of both stochastic and deterministic elements.

This initial study assumes that all forms of model uncertainty associated with climatic inputs, model parameters, and even measurement error are all embedded in the model's residual errors. Understanding how to embed such complex sources of uncertainty into ESM model residuals is an active area of research [e.g., *Clark et al*., 2008; *Schoups and Vrugt*, 2010; *Smith et al*., 2015] as is the consideration of multimodel ensemble predictions [e.g., *Clark et al*., 2015a, 2015b; *Montanari and Koutsoyiannis*, 2012].

A basic goal of this study is to document that the main problem with the deterministic use of hydrologic models is that the probability distributions of the observed and simulated responses will deviate from each other when one ignores the distributional properties of the model residuals. It has long been understood that calibrated models, either physical, empirical, or statistical, tend to produce sets of model outputs with lower variance than is experienced in the real world [e.g., *Matalas and Jacobs*, 1964; *Kirby*, 1975; *Lichty and Liscum*, 1978]. Usually, environmental simulation models are calibrated in such a way to ensure that simulated responses are unbiased, overall, when compared to the observations used to calibrate the model. Nevertheless, there is a tendency of all simulated responses to exhibit cumulative distribution functions (cdf) with flatter slopes than the observations upon which they are based. This effect has been termed the "model smoothing effect" [*Kirby*, 1975]. *Vogel* [1999] argues that, as long as model residuals are independent of model inputs, the variance of the simulated response will always be less than the variance of the observations used to calibrate the model, regardless of whether the model is based on a deterministic or stochastic representation of reality. In spite of the now widespread research, development, and application of uncertainty analysis associated with ESMs, the understanding of the uncertain properties of simulated responses has somehow failed to percolate to the level of operational hydrology. That is, designers and managers typically, though not universally, require a deterministic ("single-number") response for planning, design, and management of projects and, as is outlined here, the critical issues embedded in the development of uncertainty intervals are simply not integrated into the development of such "single-number" responses.

The divergence of the probability distributions of simulated and observed responses in deterministic hydrologic models has been recognized for decades. In the context of precipitation-runoff modeling, *Kirby* [1975] was the first to term this effect the "model smoothing effect" and to provide a simple correction useful for linear models. Numerous subsequent studies of small U.S. streams documented that rainfall-runoff model estimates of flood discharges with large recurrence intervals tend to exhibit downward bias [e.g., *Lichty and Liscum*, 1978; *Thomas*, 1982; *Sherwood*, 1994]. Additionally, *Lichty and Liscum* [1978], *Thomas* [1982], and *Sherwood* [1994] all found reductions in the variances of extreme events. Remarkably, except in instances where such models are used in forecasting, no further discussion of this issue could be found in the rainfall-runoff modeling literature. The "model smoothing effect" is not just limited to deterministic models: statistical models have also showed a reduction in simulated variance when compared with observed variance [*Skøien and Blöschl*, 2007; *Rasmussen et al*., 2008; *Farmer et al*., 2014, 2015].

For statistical models, it is often possible to derive corrections that can be used to adjust simulated responses to ensure that they exhibit the same distributional properties as the observations used to develop and calibrate the models. Within the context of regression models for extending and augmenting short streamflow records, *Matalas and Jacobs* [1964] first provided a correction to ensure the equivalence of variance between the original short record and the extended record. *Hirsch* [1979] provides a further example, which led *Hirsch* [1982] and *Vogel and Stedinger* [1985] to introduce a suite of Maintenance of Variance Extension methods that are designed to reproduce the variance of streamflows estimated from the use of regression methods. However, little attention has been paid to statistical properties beyond the variance. Simple corrections for inflating the variance of model output may be available for some particular statistical model structures, yet such corrections will be difficult to derive for more complex deterministic models and for properties other than the variance.

Interestingly, in the field of hydrologic forecasting, there is a rich history of using calibrated model residuals to correct for the "model smoothing effect" associated with both stochastic and deterministic models. For example, *Hirsch* [1981] developed a technique, known as "position analysis," for producing ensemble monthly hydrologic forecasts based on the serial correlation of monthly streamflow data. He showed that the variance of the residuals associated with an autoregressive moving-average model for streamflow forecasting could be estimated using the difference between observed monthly historical streamflow and 1-month-ahead predictions that would have been made in each preceding time step. By establishing the behavior of model residuals using a historical data set, the stochastic forecasting model could be applied for future months having, as yet, undetermined residuals. *Tasker and Dunne* [1997] document how such a stochastic "position analysis" may be applied within a drought context. More recently, *Henley et al.* [2013] document how such a stochastic "position analysis" of drought risk can be conditioned upon drivers of climate and climate change. *Clark et al.* [2004], *Schaake et al.* [2007], and others in the flood forecasting community have extended Hirsch's techniques to produce streamflow forecasts whose statistical properties mimic those upon which the models are based.

The exploration of the distributional impacts of the deterministic use of ESMs begins with a simple, statistical model, enabling a theoretical analysis that demonstrates the influence of model residuals on simulated response, highlighting the impacts of ignoring model residuals in operational hydrology. In the remainder of the paper, the properties of model output for a more realistic deterministic, distributed-parameter, precipitation-runoff model are considered. Finally, within the context of the distributed-parameter, precipitation-runoff model, an *ex post facto* solution for reincorporating model residuals into simulations and correcting for distributional bias is developed and evaluated.

2. Theoretical Properties of Simulation Model Output

This section considers some general statistical properties of simulated responses derived from a simulation model of streamflow. In general, a hydrologic model of streamflow can be abbreviated as

$$Q=\hat{Q}+\epsilon \tag{1}$$

where Q represents observed streamflow, $\hat{Q}$ represents the modeled streamflow, which is a function of model structure, input variables, and some parameter specification, and ϵ represents the model error associated with the model simulation. The examples herein address time series of streamflows and the modeling thereof, but the generalized structure of equation (1) can accommodate most spatial and temporal hydrologic models.

The first priority of model calibration is typically unbiasedness, i.e., on average the simulated response is equivalent to the observed response. This is summarized by taking the expectation of equation (1) to yield

$$E[Q]=E\left[\hat{Q}+\epsilon\right]=E\left[\hat{Q}\right]+E[\epsilon] \tag{2}$$

where $E[\ldots]$ represents the expectation operator. For the condition of unbiasedness to hold, the expectation of the residual, $E[\epsilon]$, should approach zero through calibration. While this is ideal, it is seldom achieved. Importantly, as is shown here, a model can exhibit overall unbiasedness yet still exhibit differential bias under high or low streamflow conditions or both.

Usually, the most attention during generic calibration is given to the overall variability of the residuals. Taking the variance of both sides of equation (1) leads to

$$var(Q)=var\left(\hat{Q}+\epsilon\right)=var\left(\hat{Q}\right)+var(\epsilon)+2cov\left(\hat{Q},\epsilon\right) \tag{3}$$

where $var(\ldots)$ and $cov(\ldots)$ represent the variance and covariance operators. Unlike with the expectations in equation (2), it is not as straightforward to ensure the equivalence of the variances of the simulated and observed response. There may be a complex interaction between the variance of the residuals and the covariance between the simulated response and the residuals. For simple models, it is often assumed that, by virtue of the fitting procedure and the underlying model structure, the simulated responses are uncorrelated with the residuals, that is, $cov\left(\hat{Q},\epsilon\right)=0$. If such is the case, then minimizing the variance of the residuals, $var(\epsilon)\rightarrow 0$, is the only way to ensure that the simulated responses have the same variance as the observations, $var\left(\hat{Q}\right)\rightarrow var(Q)$. Unfortunately, in practice, there is always residual error such that $var(\epsilon)>0$, thus, given the assumption of independence, the variance of the simulated responses will always be smaller than the variance of the observed responses. When the independence of the simulated responses and residuals cannot be assumed, the interplay between $var(\epsilon)$ and $2cov\left(\hat{Q},\epsilon\right)$ becomes much more important and a definitive statement concerning the relative magnitudes of $var(Q)$ and $var\left(\hat{Q}\right)$ is no longer possible.

It is not only the variance of the simulated responses that will be misrepresented when model errors are ignored: all higher and lower order moments will also be misrepresented. Consider the *i*-th moment of equation (1)

$$E\left[(Q)^i\right]=E\left[\left(\hat{Q}+\epsilon\right)^i\right] \tag{4}$$

where *i* is typically a positive integer, but, as was shown by *Farmer et al*. [2015], can be any nonzero number. *Farmer et al*. [2015] document that if one's interest is in the goodness-of-fit of a hydrologic model to the low streamflows under drought conditions, then consideration of the lower order moments, i.e., $i<0$, also termed inverse moments, is paramount. The expansion of the right-hand side of equation (4) for any positive integer documents that the equivalence of the *i*-th moments of Q and $\hat{Q}$ is dependent on the extremely complex interactions between the moments of ϵ and the moments of the cross products of ϵ and $\hat{Q}$. This dependence carries through from the noncentral moments defined above, as well as to centralized moments (e.g., variance) and standardized moments or moment ratios (e.g., the skewness and kurtosis). Furthermore, the moments of the probability distribution of streamflow are linked to the quantiles of that distribution as shown by *Farmer et al*. [2015]. Thus, and quite importantly, bias in the distributional properties of modeled streamflows will result in corresponding bias in the modeled quantiles of the streamflow distribution. However, without model-specific assumptions about the distribution of the errors (e.g., $cov\left(\hat{Q},\epsilon\right)=0$), categorical descriptions of the effects are not possible.

Due to the complex relations between the probability distributions of the observed responses, simulated responses, and model residuals, ignoring the distribution of the residuals can have a substantial impact on derived model products even beyond the characteristics of the probability distribution of the simulated response (i.e., temporal and spatial stochastic properties). The following section examines the impact of ignoring model residuals on the behavior of derived properties of model responses useful for the design, planning, and management of water resources. A simple linear model is used to arrive at first-order yet general findings, and is followed by an example using a much more realistic and complex distributed watershed model fit to hundreds of actual watersheds.

3. Example 1: A Simplified Linear Model of Streamflow

Consider a simple, analytical linear model of streamflow. As an example of a model of environmental systems, this is a gross oversimplification of reality, yet it is attractive for several reasons. The use of a simple, analytical model enables a general, closed-form demonstration of the impact of neglecting the distribution of model residuals on the output of ESMs. Such generalized results would be difficult to obtain using a realistic simulation model of a particular system. Even though the model used here is trivial, its calibration and use is representative of how most ESMs, even extremely complex and physically realistic ones, are used in practice.

Consider a precipitation-runoff model where streamflow observations, Q_t, are related to rainfall observations, P_t, by the simple linear model

$$Q_t=\alpha+\beta P_t+\epsilon_t \tag{5}$$

where α and β are model parameters, and ϵ_t are independent, identically distributed random residual errors with mean zero and a constant variance, σ_ϵ^2. An advantage of the simple model in equation (5) is that a plethora of analytical theoretical results are available. For example, *Stedinger et al*. [2008] used this model to show that the only way to obtain meaningful prediction intervals using the generalized likelihood uncertainty estimation (GLUE) method was to assume a formal statistical model for the behavior of the model residuals. Their comparisons were between well-known analytical prediction intervals for the model in equation (5) and Monte-Carlo simulation results from GLUE.

Generally, the simulation model in equation (5) would be calibrated using historical observations of precipitation and streamflow to obtain estimates of the model parameters α and β, which are denoted by *a* and *b*. After the model is calibrated, the simulated response is generated from

$$\hat{Q}_t = a + bP_t \tag{6}$$

which ignores the residual errors specified in equation (5) and used for calibration. Typically, such models are used as if they provide a deterministic representation of the underlying environmental system. As such, some modelers and many users tend to be uncomfortable considering multiple series of simulated results each of which would attempt to add the residual error back into the simulated response. This is especially true in the realm of operational hydrology, wherein managers typically desire a single deterministic answer and systems, and many existing hydrologic design methods are not equipped for handling stochastic responses. However, with the increasing popularity of uncertainty analysis, this is becoming less of a hurdle. In fact, many practitioners currently use or are moving toward the application of ensembles of similarly probable simulations. Instead of using ESMs as deterministic models, it is essential to consider similarly likely realizations of the augmented model

$$Q_t = a + bP_t + e_t$$

where the model error term, e_t, is generated in such a way as to reproduce the theoretical properties of the residual errors in equation (5).

Now consider a watershed subjected to annual precipitation, P, with mean μ_P and variance σ_P^2 resulting in streamflow, Q, with mean μ_Q and variance σ_Q^2. Given the assumption that the residuals are independent of the rainfall equation (5) one can show that the variance of the residuals is

$$\sigma_\epsilon^2 = \left(1 - R^2\right)\sigma_Q^2$$

where R^2 represents the coefficient of determination of the fitted model, a measure of goodness-of-fit, which is related to the model coefficient via

$$\beta = R\frac{\sigma_Q}{\sigma_P}$$

Simply taking the expectation of equation (5) and algebra leads to the intercept term:

$$\alpha = \mu_Q - \beta\mu_P$$

In practice, the hydrologist employs the simulated response $\hat{Q}_t$ in equation (6), whose variance is

$$var\left(\hat{Q}_t\right) = \sigma_{\hat{Q}}^2 = b^2\sigma_P^2 = R^2\sigma_Q^2 \tag{7}$$

Failing to account for the distributional properties of the residuals results in a reduction in the variance of the simulated response by a factor of R^2. According to (7), even well-fit precipitation-runoff models, which exhibit R values ranging from 0.7 to 0.9, will underestimate the variance of the observed streamflows by a factor between 0.49 and 0.81. The consequences of generating responses with variances that are too small can be tremendous, generally leading to underestimation of design flood events and overestimation of design low streamflows as is shown below. This can be especially problematic for the thousands of climate change investigations that use ESMs to explore the impacts of changes in climate inputs on environmental systems. Such models may be misrepresenting the distributional properties of their responses. (Though, as the next section will show, it is not possible to draw categorical conclusions as to the direction of bias when using more complex models.)

The underestimation of the variance of the simulated response will result in an underestimation of the design flood events and an overestimation of low-streamflow quantiles. For simplicity, consider the case of normally distributed streamflows and model residuals, in which case it is possible to derive the percent bias in resulting estimates of a quantile or design event as

$$\Delta_p Q_p = \frac{\hat{Q}_p - Q_p}{Q_p} = \frac{\left(\mu_{\hat{Q}} + Z_p\sigma_{\hat{Q}}\right) - \left(\mu_Q + Z_p\sigma_Q\right)}{\mu_Q + Z_p\sigma_Q}$$

where Q_p indicates the 100p-th percentile of the distribution of streamflow, and Z_p is the 100p-th percentile of the standard normal distribution. Since model responses are unbiased, $\mu_{\hat{Q}} = \mu_Q$, this case can be simplified by combining it with $\sigma_{\hat{Q}} = R\sigma_Q$ from equation (7) to yield

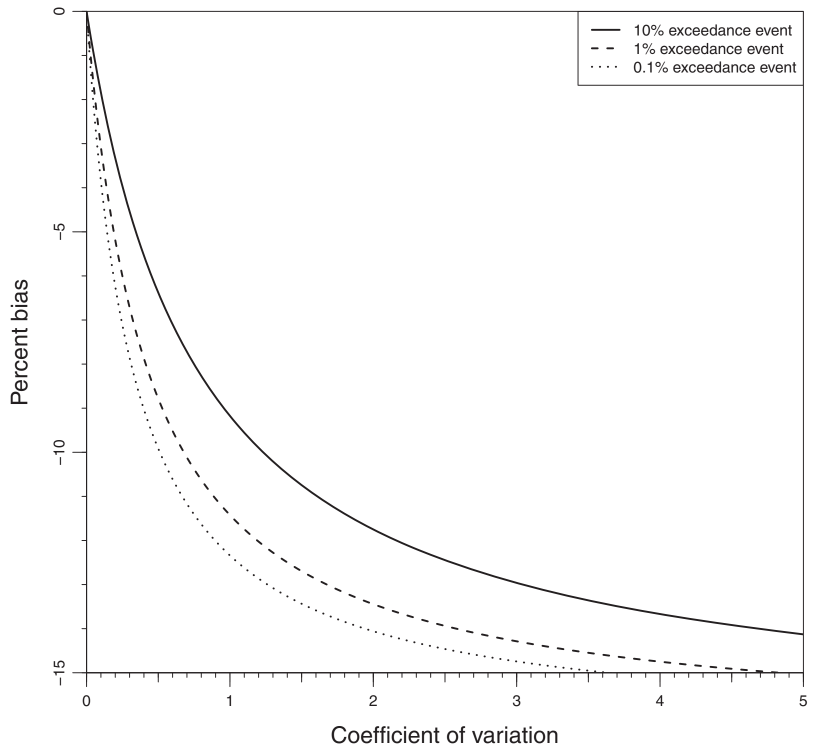

Figure 1. Example percent error in design streamflow events for a simple linear streamflow model with a coefficient of determination of $R^2 = 0.7$ ($R = 0.837$).

$$\Delta_p Q_p = \frac{\left(\mu_{\hat{Q}} + Z_p R\sigma_Q\right) - \left(\mu_Q + Z_p\sigma_Q\right)}{\mu_Q + Z_p\sigma_Q} = \frac{\left(\mu_{\hat{Q}} - \mu_Q\right) + Z_p\sigma_Q(R-1)}{\mu_Q + Z_p\sigma_Q} \tag{8}$$

Equation (8) can be used to obtain the average percent bias in design events. The T-year design event can be assessed by relating the nonexceedance probability p in equation (8) to the average return period, T, using $p = 1 - \frac{1}{T}$. Equation (8) can also be simplified by introducing the coefficient of variation, $C_Q = \frac{\sigma_Q}{\mu_Q}$ so that

$$\Delta_p Q_p = \frac{Z_p C_Q (R-1)}{1 + Z_p C_Q} \tag{9}$$

Figure 1 illustrates the downward bias associated with the design flood events in a relatively well-calibrated model ($R^2 = 0.7$). The magnitude of the bias increases as the design return period T increases and as C_Q increases. Taking the limit of equation (9) as C_Q approaches infinity reveals that the bias approaches $R-1$ (-16% in this case). Figure 1 considers downward bias associated with flood quantiles, but, due to the symmetry of the normal distribution, low flow design events will exhibit upward bias whose magnitude is identical to the downward bias shown for the high design events in this case.

4. Example 2: A Distributed-Parameter Precipitation-Runoff Model

The previous example provides a general, closed-form interpretation of the effect of ignoring the distributional properties of the residuals on simulation output and resulting design event estimates. However, several grossly simplifying assumptions were required. In real-world applications, the ESMs developed and applied are much more complex than equation (5). Accordingly, it is generally not possible to derive a closed-form solution to understand the practical implications of ignoring the distributional properties of the residuals. Instead, it is necessary to return to the theoretical underpinnings presented earlier and to observe the variation in results for a realistic model across a large number of simulation experiments.

Here, a moderately complex, distributed-parameter, precipitation-runoff model is used to observe the effect of ignoring the distributional characteristics of the residuals. The particulars of the model and the calibration scheme are not relevant to this exploration. Instead, the results serve only as an example of model output and

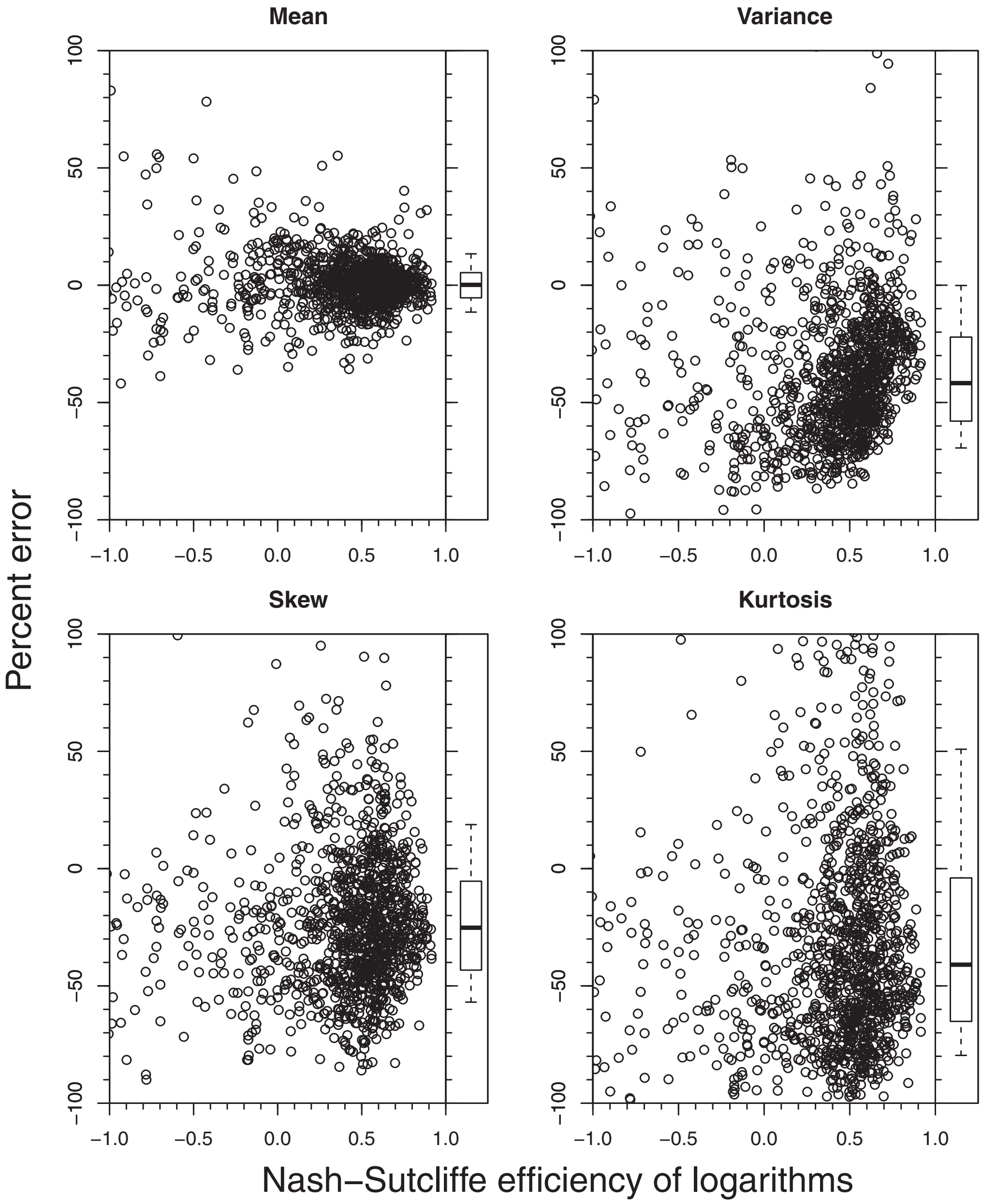

Figure 2. Percent errors in estimates of the moments, i.e., mean and variance, and the moment ratios, i.e., skewness and kurtosis, of simulated daily time series of streamflow generated from a distributed-parameter precipitation-runoff model for 1225 sites in basins across the United States Axis limits have been constrained to improve visibility while retaining 95% of plot data. The box of the box plots represents the *25th* and *75th* percentiles of the distribution of errors, while the heavy line indicates the median and the whiskers extend to the *10th* and *90th* percentiles.

the importance of accounting for the distribution of the residuals. The distributed-parameter model, in this case, the Precipitation-Runoff Modeling System [*Markstrom et al.*, 2008, 2015], was calibrated at each of 1225 perennial river basins across the conterminous United States. The focus of this work is not on the development and calibration of this model, but rather on the impacts of its deterministic use, thus further details of the model are not provided here. The general findings which follow are analogous to those of the previous case study and are in no way model dependent. The same general qualitative conclusions presented in the following section can be expected to result from the use of any hydrologic model. While this is a general calibration, and other calibration schemes may target particular extrema more directly, the theoretical argument presented remains relevant.

Figure 2 presents the percent error in estimates of the moments, i.e., mean and variance, and the moment ratios, i.e., skewness and kurtosis, of the simulated daily streamflows generated from the distributed-parameter precipitation-runoff models. The percent error is the error associated with using the simulated response rather than the observed response. The percent errors are plotted against the Nash-Sutcliffe Efficiency (NSE), a metric of general performance [*Nash and Sutcliffe*, 1970]. NSE is essentially an estimator of the standardized mean square error of the simulated responses. For many reasons, the most commonly used estimator of NSE may itself be biased and unreliable when used with daily streamflows that exhibit enormous values of skewness (see Figure 4 in *Vogel and Fennessey* [1993] and discussion). However, as the concern here is more with the impact of ignoring the model residuals, NSE is included to provide a basic frame of reference. To mitigate some of the unreliability associated with the estimator of NSE that results from the underlying skewness of daily streamflow observations, the NSE was computed on the logarithms of the daily streamflows. As expected, the models yield relatively low overall bias in the streamflows, as evidenced by the values of percent error of the mean. The percent error in estimates of the mean, also known as overall model bias, has a median of approximately zero (0.146%) across all sites. There is weak correlation between the magnitude (absolute value) of the bias and NSE, with a Spearman rank correlation of −0.327. This is not surprising, as the NSE is a measure of standardized mean square error that depends on errors in both the mean and variance.

For the majority of sites, estimates of the variance, skewness, and kurtosis of the daily streamflows are underestimated: the median percent errors are −41.7%, −25.2%, and −41.0%, respectively. Although these quantities are underestimated at the majority of sites, there are several sites (less than 25% of sites) that show a positive bias. This is due to the fact that the covariance between the simulated response and the residuals, discussed earlier with respect to equation (4), is seldom equal to zero and may be negative. The magnitude of the percent error associated with the bias of these three statistics increases with decreasing model performance. The Spearman rank correlation between NSE and those statistics are −0.275, −0.146, and −0.116, respectively. Again, one expects NSE to be inversely related to the bias associated with any streamflow statistic derived from the simulated model responses.

Estimates of the moment ratios, i.e., skewness and kurtosis, reported in Figure 2 are known to exhibit significant downward bias [*Wallis et al*., 1974], especially so for the extremely highly skewed samples of daily streamflow considered here, even those derived from very long daily flow records [*Vogel and Fennessey*, 1993]. Such severe downward bias, as documented by *Vogel and Fennessey* [1993], is inherent in all of the estimates of the moment ratios (skewness and kurtosis) reported in Figure 2 for both the observed and simulated flow series. In addition, there is also downward bias in the estimated variances, which are not moment ratios and thus not subject to the type of bias discussed by *Vogel and Fennessey* [1993]. Nevertheless, to ensure that such downward bias is not due to the issues concerning the validation of stochastic streamflow models raised by *Stedinger and Taylor* [1982], the same experiments in Figure 2 were performed using equations (11)–(14) from *Stedinger and Taylor* [1982] with the true mean assumed equal to the historical mean, as should be common practice when verifying and validating stochastic streamflow models. The results from those experiments led to similar results as reported in Figure 2 and thus are not reported here.

As was shown by *Farmer et al*. [2015] and in the previous examples, the distributional properties of the residuals and the failure to include them in an analysis will have a strong effect on estimates of streamflow quantiles. Figure 3 shows the percent error in the estimates of various percentiles of the daily streamflow distribution. Here, the exceedance probability is reported because that is common practice when illustrating flow duration curves. For each site, the integer percentiles from 1% to 99% were linearly interpolated from the observed and simulated responses. The impact of ignoring the distribution of the residuals is not as straightforward as in the previous examples. The highest streamflows are underestimated, with a median percent error of −20.4% at the exceedance probability of streamflow. The low streamflows are similarly underestimated. For example, the streamflow exceeded only 1% of the time is underestimated with a median percent error of 15.6%. There are several percentiles where the median percent error is positive. However, there are no percentiles where the direction of the bias can be uniquely categorized as positive or negative across all sites. Though not shown, there is generally a weak negative Spearman rank correlation (a median of −0.362) between the magnitude of the percent error of streamflow quantiles and Nash-Sutcliffe efficiencies of the logarithms; the magnitude of the errors increases with decreasing model performance. Figure 3 dramatizes the complexity of the impacts of ignoring model errors when employing an ESM: the resulting bias behaves in a very complex fashion across sites and across the range of streamflow exceedance probabilities of interest to hydrologists.

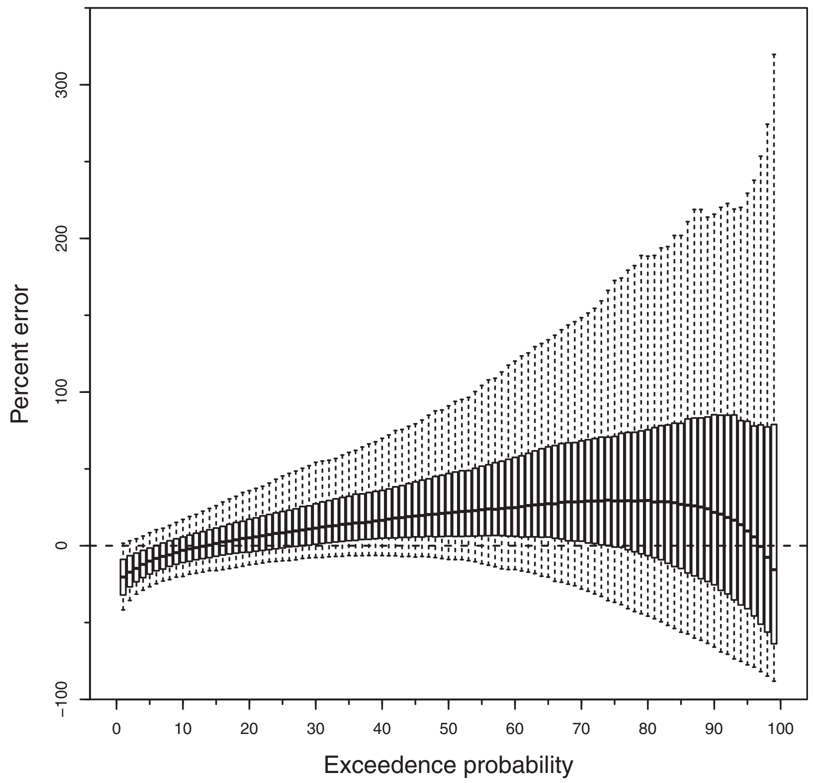

Figure 3. Percent errors in estimated quantiles of simulated daily streamflow generated from a distributed-parameter precipitation-runoff model for 1225 sites in basins across the United States Axis limits have been constrained to improve visibility while retaining 95% of plot data. The box of the box plots represents the 25*th* and 75*th* percentiles of the distribution of errors, while the heavy line indicates the median and the whiskers extend to the *10th* and *90th* percentiles.

In addition to misrepresenting the distribution of the observed streamflow response, the failure to consider the distributional properties of the residual errors can affect important decision-relevant products derived from simulated ESM responses. Figure 4 demonstrates the impact of ignoring the distributional properties of the residual error on flood frequency analysis by focusing on the series of annual maximum streamflow (flood) events derived from the model simulations. Figure 4 reports the percent errors in estimates of design flood quantiles corresponding to exceedance probabilities of 50%, 10%, 1%, and 0.1%, which correspond to estimates of the 2, 10, 100, and 1000 year quantiles of the annual maximum daily streamflow series. The median percent errors are all negative, and substantially so: −30.9%, −35.7%, −39.3%, and −41.4%, respectively. Of course, none of the events are uniformly underestimated or overestimated across all sites. Certainly, as model performance degrades, the magnitude of the errors in estimates of design events increases. The Spearman rank correlations between the daily model performance and the magnitude of the errors for the four events are −0.238, −0.230, −0.229, and −0.232. Figure 4 illustrates the remarkably large impact of ignoring model residuals on flood frequency analysis using a general calibration of a realistic precipitation-runoff model. The corresponding implications for water resources planning and management are likely to be quite considerable. Studies seeking to understand why design floods estimated from precipitation-runoff or other deterministic models do not reproduce design floods estimated from observed streamflow series [e.g., *Di Baldassarre et al.*, 2010, *Rogger et al.*, 2012, and many others] would benefit from considering the issues addressed here.

5. Reintroduction of Distributional Characteristics of Residuals to Model Simulations

The previous sections have shown the marked bias in estimates of streamflow statistics that results from ignoring the distributional properties of the residuals during model simulation. This effect can be mitigated

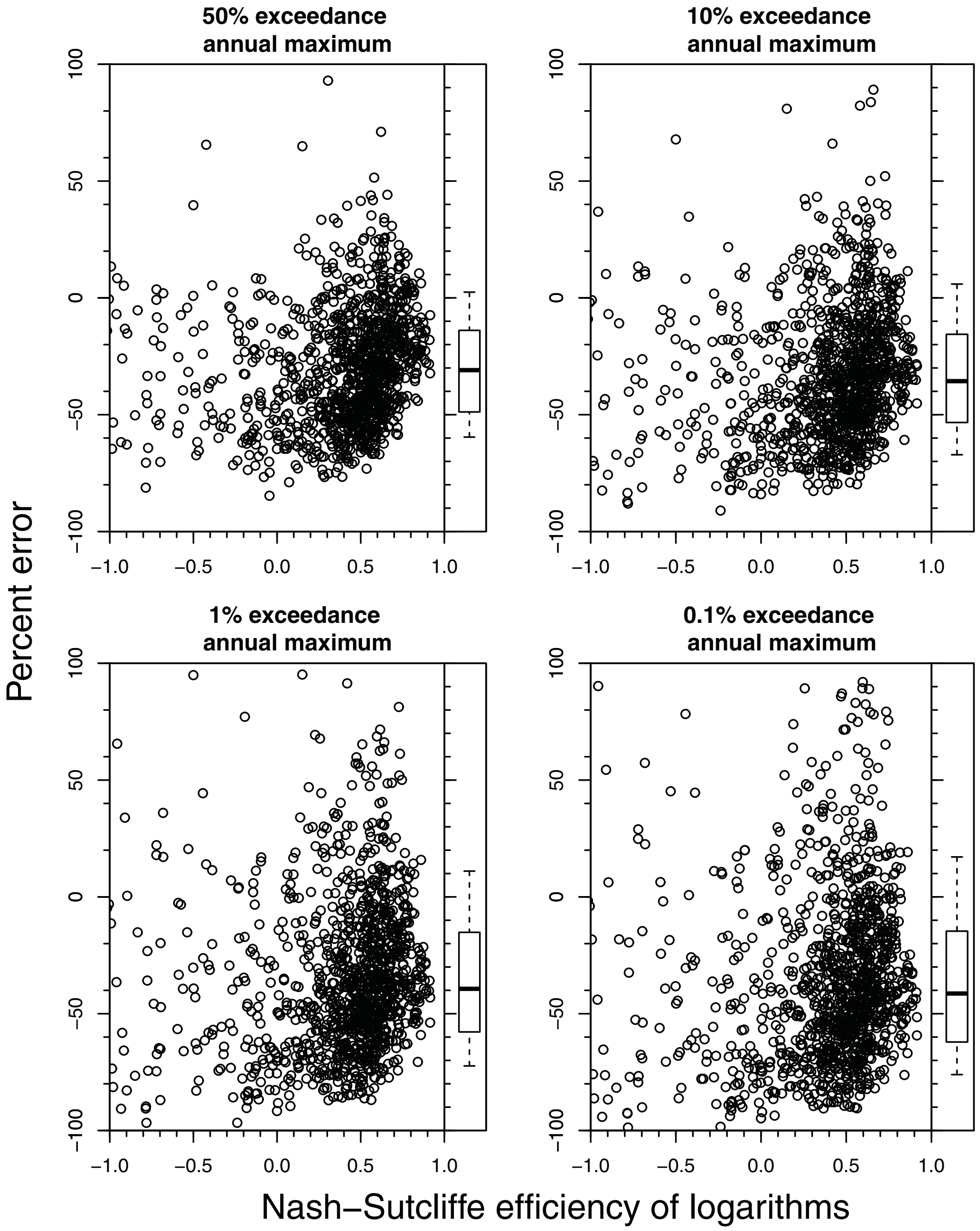

Figure 4. Percent errors in the 50%, 10% 1%, and 0.1% exceedance annual maximum streamflows from simulated daily time series of streamflow generated from a distributed-parameter precipitation-runoff model for 1225 sites in basins across the United States Axis limits have been constrained to improve visibility while retaining 95% of plot data. The box of the box plots represents the 25*th* and 75*th* percentiles of the distribution of errors, while the heavy line indicates the median and the whiskers extend to the 10*th* and 90*th* percentiles.

by attempting to reintroduce the distributional characteristics of the residuals into the simulated responses. In flood and streamflow forecasting, the Schaake Shuffle [*Clark et al*., 2004], an approach to shuffle ensemble predictions to produce new simulations preserving essential characteristics, is one method for reintroducing the distributional characteristics of model residuals into ensemble forecasts. Here, a simple resampling algorithm inspired by *Bourgin et al*. [2015] is applied. Their approach leverages the dependence between calibration residuals and predicted discharge exceedance probabilities to appropriately reintroduce model residuals. The same methodology, that of stratifying model errors sorted by exceedance probability, can be used in the at-site calibrations of the model used here.

For each site, the simulated responses were ranked and assigned a nonexceedance probability using the Weibull plotting position

$$p_i = \frac{r_i}{n+1}$$

where p_i is the nonexceedance probability of the i-th observation, r_i is the rank of the i-th observation, and n is the number of observations. The Weibull plotting position is an attractive choice because it is known to reproduce the expected value of the unknown nonexceedance probability associated with the observations, regardless of the distribution from which the observations arise [*David and Nagaraja*, 2003]. As in *Bourgin et al*. [2015], model residuals were computed as the ratio of the observed response to the simulated response. This practice assumes that the residuals are correlated with the simulated response and that each residual is simply a fraction of the original simulated response. To produce an alternate, synthetic realization of simulated responses, the simulated daily streamflows, and their associated model residuals were grouped into 10 equally spaced groups defined by the nonexceedance probabilities associated with each simulated streamflow response. For each group, the synthetic simulated response was generated by multiplying the original simulated response by a random residual drawn, with replacement, from the residuals associated with that group. Such random resampling, with replacement, is known as the Bootstrap (e.g., see drought examples in *Tasker and Dunne* [1997] and *Henley et al*. [2013]). When synthetic streamflows have been generated for all original simulations in all groups, the resulting series is presented in the original order and represents an alternate, synthetic realization of daily streamflows with model error reintroduced. This can be repeated to produce several realizations of synthetic streamflow for the same site; a process that could lead to the generation of prediction intervals, as proposed by *Bourgin et al*. [2015].

The overall approach presented by *Bourgin et al*. [2015] and modified as described above is similar to the use of copulas for implementing the Bootstrap. The stratification of residuals by the nonexceedance probability of simulated responses is, in a loose sense, a sort of nonparametric, empirical copula. However, it is not an exact copula, which would explicitly describe the joint distribution of a set of variables, because no effort is made to directly reproduce the correlation between the simulated responses and the model residuals in the residual resampling scheme. However, because the rank of the simulated responses is used to stratify the residuals, this effectively maintains the joint distribution between the residuals and the streamflows. Alternatively, as a parametric approach, it could have been assumed that simulated responses and errors follow a multivariate normal distribution. However, initial evaluations not included here but based on the assumption of multivariate normality produced synthetic series of simulated streamflows that did not mimic the observed streamflow. A different parametric copula or modified ensemble approaches [*Clark et al*., 2004; *Schefzik et al*., 2013] may be more appropriate, but our initial exploration only considers the empirical, nonparametric copula-inspired procedures described above.

Recall from Figures 2 and 3, that failing to account for the distributional properties of model residuals produced substantial errors in estimates of the various moments and quantiles of the distribution of daily streamflow. Figure 5, similar to Figure 3, illustrates the error associated with the median estimates of the various quantiles of the distribution of the synthetic daily streamflow series. Here the median errors are computed as the median of errors from 100 synthetic simulated responses produced by reintroducing model residuals as described above. Comparing Figures 3 and 5, noticing the dramatic change in vertical axis limits, there is, on average, substantial reduction in the errors associated with each quantile when model errors are reintroduced. The median reduction in error across all sites and all quantiles is 94%. For almost all quantiles, more than 90% of the sites show smaller absolute errors than were seen when ignoring the distribution of residual errors. Only at the extremes do more than 10% of sites show greater percent errors when residuals are considered. Across all sites and all quantiles, nearly 95% of site-quantiles showed improvement. The median error in the first exceedance probability is now 3.15% and the median error in the 99th exceedance probability is -18.1%; in both cases, the variability in errors is greatly reduced. Though biases remain, the improvements in the estimation of streamflow quantiles illustrated in Figure 5 have clear implications for project design. These results emphasize the need for future research into more effective stochastic methods for the reintroduction of residuals into simulation model output.

The procedure described by *Bourgin et al*. [2015] and as modified above does not account for the inherent spatial or temporal correlation structure of the model residuals. Furthermore, its application is driven only

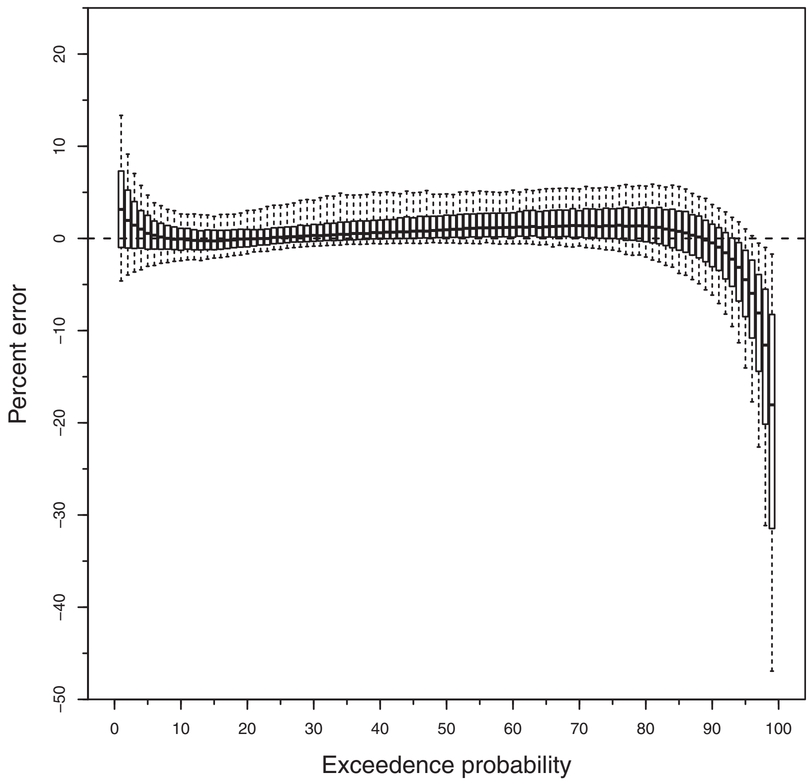

Figure 5. Median percent errors in estimated quantiles of daily streamflow generated from 100 synthetic realizations of a distributed-parameter precipitation-runoff model for each of 1225 sites in basins across the United States. Synthetic realizations were generated by reintroducing residual errors following the methods described herein. Axis limits have been constrained to improve visibility while retaining 95% of plot data. The box of the box plots represents the *25th* and *75th* percentiles of the distribution of errors, while the heavy line indicates the median and the whiskers extend to the *10th* and *90th* percentiles.

by the distribution of the simulated responses. For this reason, it is unlikely to resolve the biases associated with centralized and standardized moments or derived products such as design quantiles of the annual maximum or minimum series. The bias associated with moment estimates is complicated by the interaction and correlation among moments. The bias associated with design quantile estimates is largely due to the inherent temporal persistence in both the flow and residuals. Neither the interaction of moments nor temporal persistence are accounted for in this reintroduction method. The median percent errors of the mean, variance, skew, and kurtosis are 1.29%, 13.7%, 20.0%, and 50.8%, respectively. For the 50%, 10%, 1%, and 0.1% annual maximums, the median percent errors are 22.1% 22.3%, 22.2%, and 22.7%, respectively. Except for the kurtosis, these median percent errors show clear improvements over the previously reported values when the median of 100 realizations, which include the reintroduction of residual errors, at each site is considered. However, the improvements are highly variable and site-specific, showing that this initial method does not provide reliable performance improvements for these products.

The idea of an *ex post facto* introduction of model residuals as an improvement to simulated model responses is akin to post-processing and two-stage least squares regression. In two-stage least squares regression, an initial simulation model is developed and then a second model of the residuals or another independent variable is constructed to account for information left out of the original simulation model. A valid argument can be made that the original simulation model should just include all available information from the start. The same could be argued here but requires the redevelopment of ESMs. However, regardless of improvements in model calibration, this analysis reveals that model errors should always play a pivotal role in environmental simulation.

It is nearly impossible for the distributions of observed and simulated responses of any deterministic model to converge, exactly. Even with the reintroduction of model residuals, any single particular simulated series of responses will contain other sources of error not included in the model residuals and thus will always misrepresent the distribution of the observed response. There are many ways to calibrate a hydrologic model, but the fact remains that, even with the most targeted and evolved calibration schemes, some residual error will remain in the simulated response. The tools and discussion presented here only serve to make users aware of the danger of the deterministic use of ESMs. With this knowledge, it is possible to make more informed management decisions and judgments with regard to environmental systems and resources. In practice, the deterministic use of environmental simulation models should be tabled in favor of at least an *ex post facto* stochastic use of simulated response, whereby model residuals are reintroduced after model simulation.

6. Summary and Conclusions

The deterministic use of environmental simulation models introduces substantial distributional bias into various important statistics computed from simulated responses. This bias is particularly severe at the distributional extremes corresponding to floods and droughts, regions of particular concern to hydrologists, planners, and managers. Even though a model may generate unbiased simulations overall, bias in various important statistics derived from such simulations arise from ignoring the distributional properties of model residuals. This bias can be mitigated by considering a stochastic reintroduction of model residuals into simulated responses. Exactly how to do so requires additional exploration and may depend on the particular application and the simulation model developed. Stochastic use, as an *ex post facto* solution to the problem of distributional bias, may be a more attractive approach than attempting to improve the simulation model in question, or conducting sensitivity or uncertainty analyses. Importantly, many of the components of a rigorous uncertainty analysis can be redirected for use in the stochastic development and implementation of environmental simulation models in operational studies. Such analyses and developments may be particularly important within the context of the additional uncertainty corresponding to the impacts of climate change. The findings presented here indicate that without a consideration of the stochastic component of both statistical and deterministic models, systematic bias will result in flood and drought applications.

These conclusions are substantiated by a theoretical analysis of the impact of model residuals and further evidenced by two examples presented above. The first example considered an idealized, closed-form solution that demonstrates the general effects of ignoring the distribution of model residuals on the simulated distribution of streamflow. The second example presents the case of a more complex and realistic distributed-parameter precipitation-runoff model calibrated at each of 1225 perennial river basins across the United States There, the effects of the residuals on simulated responses cannot be explicitly derived. Instead, the effects are demonstrated by an analysis of the first four moments, the distribution of streamflow, and several design floods. The second example concludes by presenting one approach that can be used to improve the simulated response in design events. The results of this initial analysis emphasize the need to exploit recent advances in the uncertainty analysis of environmental simulation models for the purpose of developing new and effective methods of reintroducing the distributional properties of residuals into simulation output.

While this initial analysis only considered hydrologic models of streamflow time series, the general conclusions can be applied to almost any simulation modeling application in which extreme events play a pivotal role. While no single solution exists, moving from the deterministic use of environmental simulation models to at least an *ex post facto* stochastic use will improve both understanding and communication of the uncertainties and biases in simulated results, which in turn should lead to improvements in the design, planning, and management of water resource systems.

Acknowledgments
Pursuant to the U.S. Geological Survey's model and data archive policies, the data in support of Example 2 will be permanently archived with DOI:10.5066/F7W7TF4, which will be registered upon final approval and publication. Many thanks are due to Julie Kiang, Gregory McCabe, and Lauren Hay, all of the U.S. Geological Survey, who provided comments on early editions of this paper. We are also indebted to Nicholas C. Matalas and three anonymous reviewers for their constructive and insightful comments on an earlier version of the paper.

References

Beven, K. (2014), A framework for uncertainty analysis, in *Applied Uncertainty Analysis for Flood Risk Management*, edited by K. Beven and J. Hall, Imperial College Press, London, U. K.

Bourgin, F., V. Andréassian, C. Perrin, and L. Oudin (2015), Transferring global uncertainty estimates from gauged to ungauged catchments, *Hydrol. Earth Syst. Sci.*, *19*(5), 2535–2546, doi:10.5194/hess-19-2535-2015.

Clark, M., S. Gangopadhyay, L. Hay, B. Rajagopalan, and R. Wilby (2004), The Schaake Shuffle: A method for reconstructing space-time variability in forecasted precipitation and temperature fields, *J. Hydrometeorol.*, *5*(1), 243–262, doi:10.1175/1525-7541(2004)005<0243:TSSAMF>2.0.CO;2.

Clark, M. P., A. G. Slater, D. E. Rupp, R. A. Woods, J. A. Vrugt, H. V. Gupta, T. Wagener, and L. E. Hay (2008), Framework for understanding structural errors (FUSE): A modular framework to diagnose differences between hydrological models, *Water Resour. Res.*, *44*, W00B02, doi:10.1029/2007WR006735.

Clark, M. P., et al. (2015a), A unified approach for process-based hydrologic modeling: 1. Modeling concept, *Water Resour. Res.*, *51*, 2498–2514, doi:10.1002/2015WR017198.

Clark, M. P., et al. (2015b), A unified approach for process-based hydrologic modeling: 2. Model implementation, *Water Resour. Res.*, *51*, 2515–2542, doi:10.1002/2015WR017200.

David, H., and H. Nagaraja (2003), *Order Statistics*, Wiley Series in Probability and Statistics, 3rd ed., 488 pp., Wiley-Interscience, Hoboken, N. J.

Di Baldassarre, G., G. Schumann, P. D. Bates, J. E. Freer, and K. J. Beven (2010), Flood-plain mapping: A critical discussion of deterministic and probabilistic approaches, *Hydrol. Sci. J.*, *55*(3), 364–376, doi:10.1080/02626661003683389.

Farmer, W. H., S. A. Archfield, T. M. Over, L. E. Hay, J. H. LaFontaine, and J. E. Kiang (2014), A comparison of methods to predict historical daily streamflow time series in the southeastern United States, *Sci. Invest. Rep. 2014-5231*, U.S. Geol. Surv., Reston, Va., doi:10.3133/sir20145231.

Farmer, W. H., R. R. Knight, D. A. Eash, K. J. Hutchinson, M. Linhart, D. E. Christiansen, S. A. Archfield, T. M. Over, and J. E. Kiang (2015), Evaluation of statistical and rainfall-runoff models for predicting historical daily streamflow time series in the Des Moines and Iowa River watersheds, *Sci. Invest. Rep. 2015-5089*, U.S. Geol. Surv., doi:10.3133/sir20155089.

Henley, B. J., M. A. Thyer, and G. Kuczera (2013), Climate driver informed short-term drought risk evaluation, *Water Resour. Res.*, *49*, 2317–2326, doi:10.1002/wrcr.20222.

Hirsch, R. M. (1979), An evaluation of some record reconstruction techniques, *Water Resour. Res.*, *15*(6), 1781, doi:10.1029/WR015i006p01781.

Hirsch, R. M. (1981), Stochastic hydrologic model for drought management, *J. Water Resour. Plann. Manage.*, *107*(WR2), 303–313.

Hirsch, R. M. (1982), A comparison of four streamflow record extension techniques, *Water Resour. Res.*, *18*(4), 1081, doi:10.1029/WR018i004p01081.

Kirby, W. (1975), Model smoothing effect diminishes simulated flood peak variance, *Am. Geophys. Union Trans.*, *56*(6), 361.

Lichty, R. W., and F. Liscum (1978), A rainfall-runoff modeling procedure for improving estimates of t-year (annual) floods for small drainage basins, *Water Resour. Invest. Rep. 78-7*, U.S. Geol. Surv., Reston, Va.

Liu, Y. Q., and H. V. Gupta (2007), Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework, *Water Resour. Res.*, *43*, W07401, doi:10.1029/2006WR005756.

Markstrom, S. L., R. G. Niswonger, R. S. Regan, D. E. Prudic, and P. M. Barlow (2008), GSFLOW—Coupled ground-water and surface-water flow model based on the integration of the Precipitation-Runoff Modeling System (PRMS) and the Modular Ground-Water Flow Model (MODFLOW-2005), *Tech. and Methods 6-D1*, U.S. Geol. Surv., Reston, Va.

Markstrom, S. L., R. S. Regan, L. E. Hay, R. J. Viger, R. M. T. Webb, R. A. Payn, and J. H. LaFontaine (2015), PRMS-IV, the Precipitation-Runoff Modeling System, Version 4: *Tech. and Methods 6-B7*, U.S. Geol. Surv., Reston, Va.

Matalas, N. C., and B. Jacobs (1964), A correlation procedure for augmenting hydrologic data, Prof. Pap. 434, U.S. Geol. Surv., Reston, Va.

Matott, L. S., J. E. Babendreier, and S. T. Purucker (2009), Evaluating uncertainty in integrated environmental models: A review of concepts and tools, *Water Resour. Res.*, *45*, W06421, doi:10.1029/2008WR007301.

Mirzaei, M., Y. F. Huang, A. El-Shafie, and A. Shatirah (2015), Application of the generalized likelihood uncertainty estimation (GLUE) approach for assessing uncertainty in hydrological models: A review, *Stochastic Environ. Res. Risk Assess.*, *29*(5), 1265–1273, doi:10.1007/s00477-014-1000-6.

Montanari, A. (2011) Uncertainty of hydrological predictions, in *Treatise on Water Science*, edited by P. Wilderer, pp. 459–478, Elsevier, Oxford, doi:10.1016/B978-0-444-53199-5.00045-2.

Montanari, A., and D. Koutsoyiannis (2012), A blueprint for process-based modeling of uncertain hydrological systems, *Water Resour. Res.*, *48*, W09555, doi:10.1029/2011WR011412.

Moradkhani, H., and S. Sorooshian (2008), General review of rainfall-runoff modeling: Model calibration, data assimilation, and uncertainty analysis, in *Hydrological Modeling and Water Cycle, Coupling of the Atmospheric and Hydrological Models*, edited by S. Sorooshian et al., pp. 1–24, Springer, Berlin, doi:10.1007/978-3-540-77843-1_1.

Nash, J., and J. Sutcliffe (1970), River flow forecasting through conceptual models part I—A discussion of principles, *J. Hydrol.*, *10*(3), 282–290, doi:10.1016/0022-1694(70)90255-6.

Rasmussen, T. J., C. J. Lee, and A. C. Ziegler (2008), Estimation of constituent concentrations, loads, and yields in streams of Johnson County, northeast Kansas, using continuous water-quality monitoring and regression models, October 2002 through December 2006, *Sci. Invest. Rep. 2008-5014*, U.S. Geol. Surv., Reston, Va.

Rogger, M., B. Kohl, H. Pirkl, A. Viglione, J. Komma, R. Kimbauer, R. Merz, and G. Bloschl (2012), Runoff model and flood frequency statistics for design flood estimation in Austria—Do they tell a consistent story?, *J. Hydrol.*, *456*, 30–43, doi:10.1016/j.jhydrol.2012.05.068.

Schaake, J., J. Demargne, R. Hartmen, M. Mullusky, E. Welles, L. Wu, H. Herr, X. Fan, and D. J. Seo (2007), Precipitation and temperature ensemble forecasts from single-value forecasts, *Hydrol. Earth Syst. Sci. Discuss.*, *4*, 655–717, doi:10.5194/hessd-4-655-2007.

Schefzik, R., T. L. Thorarinsdottier, and T. Gneiting (2013), Uncertainty quantification in complex simulation models using ensemble copula coupling, *Stat. Sci.*, *28*(4), 616–640, doi:10.1214/13-STS443.

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

Sherwood, J. M. (1994), Estimation of peak-frequency relations, flood hydrographs, and volume-duration-frequency relations of ungaged small urban streams in Ohio, *Water Supp. Pap. 2432*, U.S. Geol. Surv., Reston, Va.

Skøien, J. O., and G. Blöschl (2007), Spatiotemporal topological kriging of runoff time series, *Water Resour. Res.*, *43*, W09419, doi:10.1029/2006WR005760.

Smith, T., L. Marshall, and A. Sharma (2015), Modeling residual hydrologic errors with Bayesian inference, *J. Hydrol.*, *528*, 29–37, doi:10.1016/j.jhydrol.2015.05.051.

Stedinger, J. R., and R. M. Taylor (1982), Synthetic streamflow generation 1: Verification and validation, *Water Resour. Res.*, *18*(4), 909–918, doi:10.1029/WR018i004p00909.

Stedinger, J. R., R. M. Vogel, S. U. Lee, and R. Batchelder (2008), Appraisal of the generalized likelihood uncertainty estimation (GLUE) method, *Water Resour. Res.*, *44*, W00B06, doi:10.1029/2008WR006822.

Tasker, G. D., and P. Dunne (1997), Bootstrap position analysis for forecasting low flow frequency, *J. Water Resour. Plann. Manage.*, *123*(6), 359–367, doi:10.1111/j.1752-1688.1982.tb03964.x.

Thomas, W. O. (1982), An evaluation of flood frequency estimates based on rainfall/runoff modeling, *J. Am. Water Resour. Assoc.*, *18*(2), 221–229, doi:10.1111/j.1752-1688.1982.tb03964.x.

Vogel, R. M. (1999), Editorial: Stochastic and deterministic world views, *J. Water Resour. Plann. Manage.*, *125*(6), 311–313, doi:10.1061/(ASCE)0733-9496(1999)125:6(311).

Vogel, R. M., and N. M. Fennessey (1993), L-moment diagrams should replace product-moment diagrams, *Water Resour. Res.*, *29*(6), 1745–1752, doi:10.1029/93WR00341.

Vogel, R. M., and J. R. Stedinger (1985), Minimum variance streamflow record augmentation procedures, *Water Resour. Res.*, *21*(5), 715–723, doi:10.1029/WR021i005p00715.

Wallis, J. R., N. C. Matalas, and J. R. Slack (1974), Just a moment, *Water Resour. Res.*, *10*(2), 211–219, doi:10.1029/WR010i002p00211.