Stochastic Residual-Error Analysis for Estimating Hydrologic Model Predictive Uncertainty

Mohamed M. Hantush, A.M. ASCE\textsuperscript{1}; and Latif Kalin, A.M. ASCE\textsuperscript{2}

Abstract: A hybrid time series-nonparametric sampling approach, referred to herein as semiparametric, is presented for the estimation of model predictive uncertainty. The methodology is a two-step procedure whereby a distributed hydrologic model is first calibrated, then followed by brute force application of time series analysis with nonparametric random generation to synthesize serially correlated model residual errors. The methodology is applied to estimate uncertainties in simulated streamflows and related flow attributes upstream from the mouth of a rapidly urbanizing watershed. Two procedures for the estimation of model output uncertainty are compared: the Gaussian-based l-step forecast and the semiparametric ensemble forecast. Results show that although both methods yielded comparable uncertainty bands, the Gaussian l-step forecast underestimated the width of the uncertainty band when compared to the semiparametric method. An ensemble of streamflows generated through Latin-hypercube Monte Carlo simulations showed relatively larger values of the coefficient of variation for long-term average annual maximum daily flows than for long-term daily, monthly maximum daily, and monthly median of daily flows. Ensemble of flow duration curves is generated from the error-adjusted simulated flows. The computed low flows displayed greater values of the coefficient of variation than flows in the medium and high range. The ensemble flow durations allow for the estimation of daily flow range upstream from the outlet with 95\% confidence for a specified design recurrence period. The computed uncertainties of the predicted watershed response and associated flow attributes provide the basis for communicating the risk to stakeholders and decision makers who are involved in the future development of the watershed.


CE Database subject headings: Watersheds; Hydrologic models; Uncertainty principles; Time series analysis; Monte Carlo method; Stochastic processes.

Introduction

Extant practices in environmental modeling are primarily concerned with model calibration and validation, but often pay much less attention to the assessment of model predictive uncertainty. The widespread use of hydrologic and water quality models in water resources management and environmental decision making, however, has raised concerns about the reliability of model predictions for decision making. Models are imperfect representations of the real world and data/observations are often incomplete or insufficient and invariably subject to uncertainty. In real world problems, uncertainties are unavoidable and should be rigorously addressed in the development and application of models. The National Research Council (NRC) (2001a) even challenged the U.S. EPA to end the practice of arbitrary selection of the margin of safety (MOS) in the total maximum daily loading program (TMDL) and promoted conducting uncertainty analysis as the basis for MOS determination. In a state-of-the-science review of the role of research in confronting the nations’ water problems, The NRC (2004) called for the explicit recognition of uncertainty occurrence, measuring its importance, and incorporating it into decision making. Imperfect model structure due to incomplete knowledge of reality, insufficient data and measurement/observation errors, spatiotemporal variability of model parameters and climate, and inadequate description of boundary conditions are all sources of uncertainty. Failure to account for these sources of uncertainty can lead to inadequate environmental decisions that can be far from satisfactory in outcome. Several research papers and reports have also emphasized uncertainty analysis as essential for hydrologic and water quality modeling (Beck 1987; Reckhow 1994a,b; NRC 2001a,b; Singh and Woolhiser 2002; NRC 2004; Refsgaard et al. 2005; Shirzohammadi et al. 2006).

While the scientific literature is replete with approaches and methods for estimating predictive uncertainty, careful examination of the hydrologic and environmental modeling literature reveals five dominant approaches including Bayesian uncertainty estimation, sampling-based methods, Pareto optimality, stochastic analysis of model residuals, and analysis of variance. Among these approaches, the sampling approach and Bayesian inferences are gaining increasing popularity in watershed model applications. The Bayesian approach recasts a deterministic model into a standard regression form and conducts model simulations based on Bayesian statistics to estimate uncertainties assuming zero-mean and normally distributed residual errors (e.g., Sorooshian and Dracup 1980; Bates and Cambell 2001; Vrugt et al. 2003b; Kavetski et al. 2006; Kuczera et al. 2006; Samanta et al. 2007; and Ajami et al. 2007). In the sampling approach, the Monte Carlo (MC) method is used to obtain the distribution of model...
output(s) by sampling model parameters from a priori probability distributions derived from literature or new knowledge gained from experience and model calibration (Binley et al. 1991; Beven and Binley 1992; Dilks et al. 1992; van der Perk and Bierkens 1997; Schulz et al. 1999; Beven and Freer 2001; Carpenter and Georgakakos 2004; Muleta and Nicklow 2005; Hossain and Anagnostou 2005; Hantush and Kalin 2005; and Arabi et al. 2007). The generalized likelihood uncertainty estimation (GLUE) methodology (Binley et al. 1991; Beven and Binley 1992; Beven 1993; Freer and Beven 1996; Schulz et al. 1999; Beven and Freer 2001; Hossain et al. 2004; Hossain and Anagnostou 2005; and Arabi et al. 2007) is one of the most widely used Bayesian-based, sampling approaches. By pooling precision-weighted information from site-specific observed data, the application of MC simulation with Bayesian inference in GLUE could lead to reduced posterior output uncertainty band and improved model parameter distributions. Rather than seeking an “optimal parameter set,” the methodology embraces the concept of “equifinality” (Beven 1993) where many different parameter combinations (behavior sets) are considered as equally likely good simulators (mimic the behavior) of the modeled system. The separation of parameter sets into behavior and nonbehavior sets in the GLUE methodology was originally introduced by Spear and Hornberger (1980). The Pareto optimality approach is another approach that is inherently deterministic and multiobjective in nature (e.g., Gupta et al. 1998; Madsen 2000; Wagener et al. 2001; Vrugt et al. 2003a); it shares a similar notion with the “equifinality” concept of GLUE methodology in the sense that the number of the (good or behavior) parameter sets, namely, Pareto sets, defines model output uncertainty. A fourth approach involves direct stochastic analysis of model residual errors (e.g., Kim and Valdés 2005; Kim et al. 2006; and Brunen and Yang 2006). In this approach, observed or residual-error time series are described by a stochastic model. The fifth strategy for quantifying uncertainty is the well-known first-/second-order analysis where the mean and variance of model output are expressed in terms of means and variances of the input random variables (Zhang and Yu 2004; Melching and Bauwens 2001; Haan 2002; and Ang and Tang 2007). While the first-/second-order approximation has a particular appeal to a wide range of engineering applications, it is limited to small parameters’ variance and requires significant coding in complex watershed models. Among these five strategies, stochastic modeling of residual errors has been given the least attention. Although this approach requires traditional manual/automated model calibration as a first step, it nonetheless applies to serially correlated non-Gaussian residual errors and accounts for both parametric and model structure uncertainties. The main premise of the methodology is that residual errors of a calibrated model are serially correlated and can be described by a quasi-parametric stochastic model.

This paper applies a hybrid time series-nonparametric probabilistic sampling approach to residual errors of a calibrated watershed model (SWAT) to estimate its predictive uncertainty. The SWAT model is calibrated and validated for daily flows in the Pocono Creek watershed which is located in eastern Pennsylvania and is experiencing rapid population growth and urbanization (Kalin and Hantush 2006a,b). The objectives of this paper are to: (1) demonstrate a novel, yet simple approach for quantifying model predictive uncertainty; (2) compare two procedures for the estimation of uncertainty band, namely, the Gaussian-based l-step forecast and the semiparametric method; (3) evaluate model forecast performance; and (4) compute ensemble of duration curves and estimate long-term statistics of flow attributes that have management and ecological significance.

The remainder of the paper is organized in sections and subsections. It devotes an entire section to the development of a time series model for model residual errors, construction of the l-step forecast, and description of the semiparametric approach. The subsequent section deals with the application of the methodology to the calibrated SWAT model for the Pocono Creek watershed, followed by discussion of the results. The paper ends with the conclusions.

Methodology

Prediction Error

We start by noting that while model uncertainty is actually the deviation of a calculated value from the true value, the analysis that follows rather focuses on the deviation of calculated value from the observation. Model error, \( e_t \), is thus defined by this equation

\[ e_t = o_t - q_t, \quad t = 1, 2, \ldots, T \]

where \( o_t \) = measured streamflow; \( q_t \) = model computed streamflow; \( t \) = time index (days); and \( T \) = length of the time record. The \( e_t \) accounts for model structural uncertainty, parametric uncertainty, measurement error, and errors in the precipitation and other model inputs (e.g., surface elevation, climatic data, channel morphology). It should be noted that an attempt to predict \( e_t \) will yield a forecast for streamflow excluding the effect of measurement errors. The smaller the measurement error, the closer the forecast is to actual flow values.

Fig. 1(a) shows the time series of residual errors \( e_t \) of a SWAT hydrologic model calibrated and validated by Kalin and Hantush (2006b) for the Pocono Creek watershed for the period July 1, 2002–April 30, 2005. The 120 km² Pocono Creek watershed is located between the latitudes 40°59’N–41°06’N and longitudes 75°14’W–75°26’W in Monroe County, Pa., situated within the Delaware River Basin. The watershed valley is 26 km long and drains from the Pocono Plateau in its headwaters eventually into the Brodhead Creek which is a tributary to the Delaware River. The model is constructed for the area above a United States Geological Survey (USGS) stream gauge located about 6.4 km upstream from the mouth near the city of Stroudsburg, Pa., and drains an area of about 98.87 km². The model was calibrated and validated using a split data set approach and the Nash–Sutcliffe efficiency for the simulated outlet daily streamflows used in this analysis was 0.74 for the calibration and 0.71 for the validation runs. The coefficient of determination (\( R^2 \)) was 0.74 and 0.72 for the calibration and validation runs, respectively (Kalin and Hantush 2006b). These performance evaluation values are well above the threshold values recommended by Santhi et al. (2001) for SWAT calibration. However, both performance evaluation measures were relatively low for the simulated base flow, but with less than 5% mass balance error at the monthly time scale.

A close look at the daily \( e_t \) time series reveals systematic errors shown as large positive or negative spikes [Fig. 1(a)] that are symptomatic of SWAT performance at the daily time scale and appears to be not random in nature. In many instances a large (–) error is immediately followed by a large (+) error. One reason might be timing errors in recording either the precipitation or the streamflow. Another, probably more reasonable, explanation is the inability of the SWAT model to use subdaily rainfall data. When a
large rainfall event happens at the very end of a day, the actual watershed response will be about 1 day delayed compared to what SWAT actually simulates with daily rainfall data. This will clearly result in overprediction (negative error) and underprediction (positive error) on 2 successive days, respectively. Hence, to minimize such types of errors to some extent, we decided to use the average over 3 consecutive days (i.e., 3-day moving average) as a surrogate to daily flows. Two compelling reasons drove the need for consideration of simulations at the daily scale. First, monthly median daily flow is a required flow metric (or attribute) for wild trout habitat condition, as one of the primary objectives of the calibrated watershed model is to assess the impact of projected future buildout in the watershed on the sustainability of the wild brown trout population in the Pocono Creek. Second, there is not sufficiently long monthly streamflow data to conduct a meaningful time series analysis at the monthly time scale. Therefore, in addition to smoothing the effect of successive large (−) and (+) errors during some significant events, 3-day averaged streamflows provide surrogates to daily flows for use in the wild trout habitat assessment analysis. Later in the analysis, it will be shown through regression relationships derived from model simulated daily flows that estimates of 3-day averaged flow attributes can be converted back to the daily ones with high efficiency ($R^2 > 0.97$).

**Time Series ARIMA Model Development**

Fig. 1(b) shows the 3-day average of model residual errors for the calibration and validation period (July 1, 2002–April 30, 2005). The errors are not totally eliminated but their magnitudes are significantly reduced. The Minitab statistical computer package is used to identify a time series model and construct a reasonable forecast for future $e_t$ values. In the most general form, the sequence $e_t$ may be described by a multiplicative seasonal autoregressive integrated moving average model denoted by ARIMA $(p,d,q) \times (P,D,Q)_S$. The parameters $p$, $d$, and $q$ are, respectively, the orders of the autoregressive, difference, and moving average operators; $P$, $D$, and $Q$ are, respectively, the orders of the seasonal autoregressive, seasonal difference, and seasonal moving average operators; and $S$ is the seasonal period. Autocorrelation function (ACF) and partial autocorrelation function (PACF) are utilized and the procedure outlined by Shumway (1988), which derives its basis from Box and Jenkins (1970), is followed to identify the best model. The objective of the model identification process is to produce identically distributed, independent residuals $w_t$, with a minimum variance $\sigma_w^2$ arising from fitting an ARIMA $(p,d,q) \times (P,D,Q)_S$ to the $e_t$ time series.

Figure 2(a) shows the computed computed ACF and PACF of $e_t$. It is obvious that the series display nonstationarity accentuated by the slowly decaying ACF as a function of lag and the large positive value of PACF at Lag 1. The ACF of the first difference in Fig. 2(b) contains a fairly strongly negative peak (−0.44) at Lag 3 and zero thereafter, indicating that a seasonal moving average with $S=3$ (days) and $Q=1$ might be appropriate. The decreasing peaks of the PACF at multiples of three are due to the seasonal moving average component. Therefore, ARIMA $(0,1,0) \times (0,0,1)_3$ appears to be, at the moment, a suitable selection among, perhaps, other competing models. We experimented by adding autoregressive and moving average components progressively. Table 1 lists some of the examined ARIMA models, with computed variance of residuals, $\sigma^2_w$, and various goodness-of-fit measures: final prediction error (FPE), Akaike’s information criterion (AIC), and Bayesian information criterion (BIC). It is clear that the most appropriate model choice may be ARIMA $(1,1,1) \times (0,0,1)_3$, which has the smallest values of $\sigma^2_w$, FPE, AIC, and BIC. The model was fitted to the error time series depicted in Fig. 1(b) (recall the 3-day averages of the actual model errors), and the equation describing ARIMA $(1,1,1) \times (0,0,1)_3$ is...
\[
(1 - \phi B) \nabla e_i = (1 - \Theta B^3)(1 - \Theta B)w_i
\]

where \( \nabla e_i = e_i - e_{i-1} \) = first difference operator; \( B e_i = e_{i-1} \) = backward shift operator; and \( \phi, \theta, \text{ and } \Theta = \) fitted parameters. Equation (2) can be expanded to yield the following representation for \( e_i \):

\[
e_i = (1 + \phi)e_{i-1} - \phi e_{i-2} + \psi w_i - \theta w_{i-1} + \theta \Theta w_{i-4}
\]

\( \phi, \theta, \text{ and } \Theta \) are estimated by maximizing the log likelihood function for \( w_i \), \( t=1, \ldots, T \), with maximum likelihood estimates of \( \hat{\phi} = 0.78 \), \( \hat{\theta} = 0.62 \), and \( \hat{\Theta} = 0.98 \), respectively. The maximum likelihood estimate for the variance \( \sigma_w^2 \) is \( \hat{\sigma}_w^2 = 0.629 \). The ACF and PACF of \( w_i \) from this model is plotted in Fig. 2(c), and shows no prominent peaks at 5% significance limits. Therefore, it seems reasonable to regard the residuals as being white noise. The Box–Pierce \( Q \) for Lags 48, 100, and 250 measuring randomness are satisfied with 95% confidence for the selected ARIMA \( (1,1,1) \times (0,0,1)_3 \) model.

In the following sections, we present two methods for the estimation of uncertainty limits of the simulated streamflows. The first is referred to as the \textit{l-step forecast} and the second as the \textit{semiparametric} method.

**I-Step Forecast**

The \textit{l-step forecast} is given by the conditional mean (Shumway 1988)

\[
e'_t = E[e_t | e_{t-1}, \ldots]
\]

and the forecast variance

\[
P'_t = E[(e'_t - e'_t)^2 | e_{t-1}, \ldots] = \sigma_w^2 \sum_{k=0}^{l-1} \Psi_k^2
\]

where \( \Psi_k \) is to be determined later by noting that \( e_t \) can be expressed in terms of \( w_i \).
These values can be approximated by assuming that \( w_2 = w_1 = w_0 = w_{-1} = 0 \) and rewriting Eq. (3).

\[
w_2 = \theta w_{k-1} + \Theta w_{k-3} - \Theta w_{k-4} + \epsilon_k - (1 + \phi) \epsilon_{k-1} + \phi \epsilon_{k-2}
\]  

(14)

for \( k = 3, 4, \ldots, t \).

The coefficients \( \Psi_1, \Psi_2, \ldots \) in Eq. (6) can be found explicitly by substituting Eq. (6) into Eq. (2) and equating coefficients of \( B^i \), \( i = 1, 2, 3, \ldots \) in the relation

\[
(1 - \phi B)(1 - B)(1 + \Psi_1 B + \Psi_2 B^2 + \ldots) = (1 - \Theta B^3)(1 - \theta B)
\]  

(15)

This leads to

\[
\Psi_1 = 1 + \phi - \theta
\]  

(16)

\[
\Psi_2 = (1 + \phi) \Psi_1 - \phi
\]  

(17)

\[
\Psi_3 = (1 + \phi) \Psi_2 - \phi \Psi_1 - \Theta
\]  

(18)

The recursive relationships Eqs. (16)–(20) can be substituted into Eq. (5) to compute the \( l \)-step forecast variance as the mean-square error of prediction. For normally distributed \( w_t \), the \((1-\alpha)\) probability interval (confidence band) for the residual forecast value is

\[
e_{f,t} \pm z_{\alpha/2} \sqrt{\hat{\sigma}_{\hat{w}}^2}
\]  

(21)

where \( z_{\alpha/2} \) is upper \( \alpha/2 \) tail on the standard normal distribution. Although not shown here, plotting positions reveal that the computed residuals \( \hat{w}_t \) are non-Gaussian. While this violates the Box–Pierce goodness of fit, which is based on normally distributed \( w_t \), the ACF and PACF in Fig. 2, \( \hat{\sigma}_{\hat{w}}^2 \), and comparison of FPE, AIC, and BIC criteria altogether point toward a suitable model fit. For non-Gaussian \( w_t \), the \( l \)-step forecast Eqs. (9)–(13) and its variance Eq. (5) are no longer sufficient to infer the \((1-\alpha)\) confidence band; however, a nonparametric method can be employed to construct an empirical probability density function for \( w_t \).

**Table 1. Values of Residual Variance (\( \sigma_{\hat{w}}^2 \)), Final Prediction Error (FPE), Akaike’s Information Criterion (AIC), and Bayesian Information Criterion (BIC) for Various Models Applied to Model Prediction Error Data, \( e_t \).**

<table>
<thead>
<tr>
<th>ARIMA ((p,d,q)\times(P,D,Q)_I) model</th>
<th>( \hat{\sigma}_{\hat{w}}^2 )</th>
<th>FPE</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>((0,1,0) \times (1,0,0)_I)</td>
<td>0.926</td>
<td>0.928</td>
<td>-0.075</td>
<td>-0.070</td>
</tr>
<tr>
<td>((0,1,0) \times (2,0,0)_I)</td>
<td>0.805</td>
<td>0.805</td>
<td>-0.213</td>
<td>-0.203</td>
</tr>
<tr>
<td>((0,1,0) \times (0,0,1)_I)</td>
<td>0.674</td>
<td>0.675</td>
<td>-0.393</td>
<td>-0.388</td>
</tr>
<tr>
<td>((0,1,1) \times (0,0,1)_I)</td>
<td>0.653</td>
<td>0.656</td>
<td>-0.422</td>
<td>-0.413</td>
</tr>
<tr>
<td>((0,1,2) \times (0,0,1)_I)</td>
<td>0.641</td>
<td>0.645</td>
<td>-0.439</td>
<td>-0.425</td>
</tr>
<tr>
<td>((1,1,0) \times (0,0,1)_I)</td>
<td>0.647</td>
<td>0.650</td>
<td>-0.432</td>
<td>-0.422</td>
</tr>
<tr>
<td>((1,1,1) \times (0,0,1)_I)</td>
<td>0.629</td>
<td>0.633</td>
<td>-0.458</td>
<td>-0.443</td>
</tr>
</tbody>
</table>

**Semiparametric Method**

For \( w_t \) with arbitrary PDF, we use the parametric relationship Eq. (3) as a base for constructing the SWAT prediction uncertainty band. When commonly used parametric probability distributions (e.g., normal, log-normal, exponential, etc.) poorly fit the frequency of the observed stochastic series, the probability density function of the observed random variable can be locally approximated by a nonparametric model (Lall 1995). We therefore explore a nonparametric model fit to the observed probability density functional (PDF) of \( w_t \).

In parametric methods, the density function is estimated by assuming that data are drawn from a known parametric family of distributions. The methods of moments, maximum likelihood estimation, or any other methods are commonly used to estimate the parameters of the chosen PDF. In nonparametric approaches, a kernel function is often used to generalize the density function estimation. Given a set of \( n \) observations \( w_t, w_2, \ldots, w_p \), a mathematical expression of a univariate kernel probability density estimator is (e.g., Kim and Valdés 2005)

\[
\hat{f}_K(x) = \frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{x - w_k}{h}\right)
\]  

(22)

where \( x \) is given value of the random variable \( X \) denoting \( w_t \); \( K \) is kernel function; \( n \) is number of observations; and \( h \) is bandwidth that controls the variance of the kernel function. Kim and Valdés (2005) provide a list of kernel functions typically used in hydrology; the most widely used one being the Gaussian

\[
K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}
\]  

(23)

An optimal estimate for the bandwidth for a Gaussian kernel is provided by Kim and Valdés (2005) citing Silverman (1986)

\[
\hat{h} = \left(\frac{4}{3n}\right)^{1/5} \sigma
\]  

(24)

where \( \sigma \) is standard deviation of the observed record, which in this case is equal to the standard deviation of observed residuals.
that future improved local fit with a longer tail. However, one should bear in mind that a nonparametric approach is limited by the hypothesis that future \( w_t \) has a similar nonparametric functional form of the fitted PDF. Both distributions appear to be symmetric with almost zero mean and medians.

Random sampling of \( w_t \) is achieved in two steps. First, the cumulative distribution function (CDF), \( F_X(x) \), is constructed by integrating Eq. (22) (with \( u \) replacing \( x \) in the integrand) from \( u = -\infty \) to \( u = x \) to yield

\[
F_X(x) = \frac{1}{2} + \frac{1}{2n} \sum_{k=1}^{n} \text{erf} \left( \frac{x - w_k}{\sqrt{2}h} \right)
\]

where \( \text{erf}(u) = \frac{2}{\sqrt{\pi}} \int_{0}^{u} e^{-y^2} dy \) is well known error function. Second, a procedure for sampling \( x \) from \( F_X(x) \) starts with generating a random number \( R \) from the uniform distribution in the interval \([0, 1]\) (e.g., Haan 2002), then setting \( R = F_X(x) \) and solving for \( x = F_X^{-1}(R) \) using any of the finding searching techniques.

A large sequence of independent, identically distributed model residuals \( w_t \) are generated randomly using the foregoing procedure and subsequently fed into Eq. (3) to synthesize an ensemble of \( e_t \) time series. Latin-hypercube sampling (LHS) is employed as an efficient and effective alternative to conventional Monte Carlo sampling (MCS). The LHS (McKay et al. 1979) divides the CDF (i.e., the \( F_X(x) \) function) into segments of equal width from each of which a random variate \( R \) is generated. This way the whole CDF is covered, but with a smaller number of replications than the MCS. LHS is more efficient than MCS at both ends of the CDF. Finally, the streamflow forecast is obtained by adding each of the simulated daily streamflow rates.

Application

Model Forecast Evaluation

To examine model forecast performance, the period from July 1, 2002 to April 30, 2005 (1,035 days) is divided into two parts. The first 900 days (July 1, 2002–December 16, 2004) were used to construct the residual error time series model and the remaining 135 days from (December 16, 2004–April 30, 2005) were used for validation. The (1-\( \alpha \)) confidence 1-step forecast is obtained by combining the SWAT simulation and \( e_t \) forecast

\[
\hat{q}_{t+1} = q_{rel} + e_{rel} + \hat{e}_t
\]

where \( q_{rel} \) is estimate of 3-day average of the streamflow; \( e_{rel} \) = sampled sequence of \( e_t \). For each \( t \) (i.e., day), the median and uncertainty band are obtained from the corresponding ensemble of 500 values of \( \hat{q}_t \). As pointed out earlier, the uncertainty band computed with Eqs. (26) and (27) actually provides a degree of confidence in the computed value relative to the observation rather than the true value.

Figs. 4 and 5 compare observed streamflow during the validation period with the 95% confidence band and the median of \( \hat{q}_t \) computed by the two methods. The 1-step forecast with an assumed Gaussian white noise \( (w_t) \) produced a relatively narrower confidence band (Fig. 4) than that constructed based on the semiparametric model, with nearly twice the number of observations falling outside the 95% confidence band for the former method. About 7% of the observed values fall outside the 95% confidence band for the semiparametric method. Although not shown at the depicted scale, two large values of the observed streamflows fell way outside the confidence band during the time period March 26–April 15, 2005 for both methods. However, the bulk of the observed streamflows fell within the uncertainty band for both methods. The median of the 1-step forecast and the semiparametric method Compared fairly well with the measurements.

To further examine the model forecast during low-flow conditions, the forecast evaluation period is extended from May 1, 2005 to September 30, 2005. This period of data was not available during the initial model calibration/validation efforts by Kalin and Hantush (2006b). During this period, no significant storm events were recorded. Although the simulated median daily flows generally compared well with measured counterparts further validating the calibrated model, the 95% confidence bands, however, were wide for both the 1-step forecast and semiparametric ensemble forecast (Fig. 6); the former still underestimates the uncertainty bandwidth.

Ensemble Streamflow Attributes

In this section and the following one, we estimate distributional characteristics of key streamflow attributes that have management and ecological significance, and construct ensemble streamflow duration curves. We use SWAT and the stochastic residual-error model as a base to generate an ensemble of 500 20-year long time series of daily streamflows (actually 3-day averages) through MC simulations. Table 2 lists the mean, standard deviation (SD), co-
efficient of variation (COV), median, and 2.5th and 97.5th percentiles (95% confidence interval) for average daily flow, average monthly median of daily flows, average monthly maximum daily flow, and average annual maximum daily flow, each of which is computed from the ensemble of 500 time series of the synthesized daily streamflows.

In obtaining the ensemble of 20-year streamflows, rather than using historical data and in order to account for uncertainty in future precipitation values, we used the SWAT built-in weather generator WXGEN to generate 500 sets of 20-year long records of daily precipitation at the two gauge stations that were used in the model calibration (see Kalin and Hantush 2006a). Each of the synthesized precipitation sequences is fed into SWAT to obtain a 20-year long time series of daily streamflows. A 30 year warmup period is used to eliminate the effect of unknown initial conditions. For each MC simulation, 3-day averages are computed from the simulated daily streamflows. Similarly, an ensemble of 500 20-year long sequences of $t$ is generated as outlined above, and each sequence (realization) is superimposed on each SWAT MC simulation to obtain an ensemble forecast of streamflows at the USGS stream gauge, 6.4 km upstream from the mouth of the watershed.

In Table 2 summary statistics related to high-flow conditions as well as long term averages are given. This method did not perform well for low-flow index, e.g., $7Q_{10}$, because generated large negative errors superimposed on streamflows resulted in negative flow values (about 12% of all the synthesized flows were negative). Therefore, low-flow indices were not computed. The term “average” in each column title denotes the arithmetic average over the 20-year simulation period. For each realization (i.e., time series) out of 500, daily flows are averaged over the 20-year simulation record to yield “average daily flow.” Thus, there are 500 such “average daily flow” values from which the mean, median, COV, and 95% confidence limits were computed. The average monthly median of simulated daily streamflow in column two represents the arithmetic average over the simulation period of the monthly median values. As indicated earlier, this is a required metric to a fish habitat model used by the Pennsylvania Fish and Boat Commission. In column three the monthly maximum daily streamflow averaged over the simulation period are given. Column four shows the annual maximum daily streamflows averaged over the simulation period. The average daily flow, average monthly median of daily flows, and average monthly maximum daily flow have small COV values and relatively narrower 95% confidence bands. The average annual maximum daily flow, on the other hand, shows a relatively greater uncertainty, as reflected by the larger COV and wider 95% confidence band. Note that the mean and median of the four streamflow characteristics are almost equal, indicating symmetric distributions. Compared to absolute predictions of daily

Fig. 4. 1-step forecast (assuming normality of $w_t$) of 3-day averaged streamflows and measured counterparts during validation period (December 16, 2004–April 30, 2005). Total number of observed flows that are outside 95% confidence band is 18 (13.3%).

Fig. 5. Semiparametric ensemble forecast (median and 95% confidence band) and measured streamflows during validation period (December 16, 2004–April 30, 2005). Total of nine measurements lie outside 95% confidence band (6.7%). Simulated and measured values are 3-day averages.
streamflows—which can be highly uncertain as shown in Fig. 6—these flow attributes appear to be fairly predictable and therefore more reliable and suitable for design purposes.

**Ensemble Flow-Duration Curves**

A flow duration curve (FDC) is constructed for each time series \( \hat{q}_t \) in the ensemble forecast. Fig. 7(a) plots the median of the computed FDCs for the 3-day averaged streamflow rates and the 95% confidence interval. A FDC is a plot of the flow rate \( x \) versus probability of exceedance, \( P(X > x) \). The return period, \( T_r \), defined as the average recurrence interval between events producing flow rates equal to or greater than a specified flow magnitude, \( x \), is the reciprocal of \( P(X > x) \): 
\[
T_r = \frac{1}{P(X > x)}
\]
High flows typically are associated with small probability of exceedance, whereas low flows are associated with large probability of exceedance or, equivalently, with small return period. The ensemble of FDCs can be interpreted as follows. As an example, from Fig. 7(a) it can be seen that for \( P(X > x) = 0.002 \), daily streamflow at the gauge station is between 16 and 25 m\(^3\)/s with 95% confidence. Since a return period \( T_r = 1/0.002 = 500 \) days corresponds to 0.002 probability of exceedance, then with 95% confidence, the \( T_r = 500 \) day is between 16 and 25 m\(^3\)/s. Similarly, from the fig-

| Table 2. Summary of Ensemble Statistical Parameters of Selected Streamflow Characteristics Obtained from 500 Time Series of Synthesized 3-Day Average Streamflows; Values within Parentheses Denote Daily Statistics Converted from 3-Day Statistics Using Regression Relationships Shown in Fig. 9 |

<table>
<thead>
<tr>
<th></th>
<th>Average daily flow (m(^3)/s)</th>
<th>Average monthly median daily flow (m(^3)/s)</th>
<th>Average monthly maximum daily flow (m(^3)/s)</th>
<th>Average annual maximum daily flow (m(^3)/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>2.42</td>
<td>2.10</td>
<td>7.12</td>
<td>17.39</td>
</tr>
<tr>
<td>SD</td>
<td>0.203</td>
<td>(2.06)</td>
<td>(8.06)</td>
<td>(20.27)</td>
</tr>
<tr>
<td>COV</td>
<td>0.084</td>
<td>0.182</td>
<td>0.411</td>
<td>1.93</td>
</tr>
<tr>
<td>Median</td>
<td>2.42</td>
<td>2.10</td>
<td>7.13</td>
<td>17.43</td>
</tr>
<tr>
<td>95% C.I.</td>
<td>2.02–2.80</td>
<td>1.75–2.45</td>
<td>6.35–7.91</td>
<td>13.91–21.82</td>
</tr>
</tbody>
</table>

*Fig. 6. SWAT model forecast (median and 95% confidence band) and measured streamflows during postvalidation period (May 1, 2005–September 30, 2005): (a) 1-step forecast; (b) semiparametric ensemble forecast. Simulated and measured values are 3-day averages.*
note that computed streamflow rates that vary from about uncertainty than medium range and high flows. It is interesting to
greater COV values for low flows and, thus, shows much higher
of uncertainty. Close inspection of Fig. 8, however, reveals
probability of exceedance can be visually illusive as to the degree
between the transformed quantities. The log transformation
yielded magnified synthesized flows and the Box–Cox transform-
mation was limited by the condition of having a positive base in
the power term. Thus, analyses are carried out without transform-
However, the smoothing produced by the 3-day moving
average resulted in reasonably synthesized surrogates to the simu-
ulated daily streamflows, as bulk of the observed 3-day averaged
streamflows fell well within the constructed 95% confidence
bands for the validation and postvalidation periods. This is not-
withstanding the relatively wide confidence band obtained for the
validation period (Fig. 6) during which no significant rainfall
events have occurred. The 3-day average flows used in the analy-
ses in fact could be considered as transformation of daily flows.
Fig. 8 provides linear regression relationships, which are obtained
model simulation results, to convert streamflow attributes of
3-day averages back to daily ones. Values shown in parentheses
in Table 2 are daily flow statistics obtained through regression
relationships from measured streamflows, yet the coexistence of negative and positive values prevented the
commonly used log (i.e., ln εs) and Box–Cox transformations
\([((εs+1)^{−1})/λ]\). We also tried transforming observed and model
simulated flows, and defined the model error as the difference
between the transformed quantities. The log transformation
yielded magnified synthesized flows and the Box–Cox transfor-
mation was limited by the condition of having a positive base in
the power term. Thus, analyses are carried out without transform-
However, the smoothing produced by the 3-day moving
average resulted in reasonably synthesized surrogates to the simu-
lated daily streamflows, as bulk of the observed 3-day averaged
streamflows fell well within the constructed 95% confidence
bands for the validation and postvalidation periods. This is not-
withstanding the relatively wide confidence band obtained for the
validation period (Fig. 6) during which no significant rainfall
events have occurred. The 3-day average flows used in the analy-
yses in fact could be considered as transformation of daily flows.
Fig. 8 provides linear regression relationships, which are obtained
model simulation results, to convert streamflow attributes of
3-day averages back to daily ones. Values shown in parentheses
in Table 2 are daily flow statistics obtained through regression
relationships shown in Fig. 9. Similar regression relationships
could be obtained for other flow attributes and their statistical
properties. For instance, the FDC for the daily flow could be
obtained from the 3-day average FDC as follows:
1. Divide the FDC into segments by using the exceedance
probabilities as the breakpoints;
2. From the model simulations compute the flows \((Q)\) corre-
sponding to each exceedance probability \((p)\), \(Q_p\), for daily
and 3-day average flows. Repeat this for each MC simulation
result;
3. Regress \(Q_p\) values obtained from daily and 3-day average
flows for each break point \((p)\); and
4. Use the regression equations obtained in step (3) as a base to
convert the synthesized 3-day average flows to daily flows.
By following the above procedure we converted the ensemble
of 3-day average FDC shown in Fig. 7(a) to the ensemble of daily
FDC [Fig. 7(b)]. The exceedance probabilities of 0.001, 0.01,
0.05, 0.1, 0.2, . . . , 0.8, 0.9, 0.95, 0.99, 0.999 are used as the
breakpoints, and regression equations are generated at these loca-
tions with \(R^2=0.93\). Of course, the ideal procedure is to construct
such regression relationships from measured streamflows,
provided that adequately long historical records are available.
Unfortunately, this is not the case here. Comparison of the two

Discussion
The presented analysis was preceded by an attempt to fit the
ARIMA model to the daily residual errors. The shear number of
spikes in the residual errors, which are inherent to the SWAT
model during storm events, resulted in numerous synthesized
negative streamflow values, thus precluded any meaningful uncer-
tainty analysis. We attempted to transform the model errors, \(e_s\),
yet the coexistence of negative and positive values prevented the
commonly used log (i.e., ln \(e_s\)) and Box–Cox transformations
\([((e_s+1)^{−1})/λ]\). We also tried transforming observed and model
simulated flows, and defined the model error as the difference
between the transformed quantities. The log transformation
yielded magnified synthesized flows and the Box–Cox transform-
mation was limited by the condition of having a positive base in
the power term. Thus, analyses are carried out without transform-
However, the smoothing produced by the 3-day moving
average resulted in reasonably synthesized surrogates to the simu-
lated daily streamflows, as bulk of the observed 3-day averaged
streamflows fell well within the constructed 95% confidence
bands for the validation and postvalidation periods. This is not-
withstanding the relatively wide confidence band obtained for the
validation period (Fig. 6) during which no significant rainfall
events have occurred. The 3-day average flows used in the analy-
yses in fact could be considered as transformation of daily flows.
Fig. 8 provides linear regression relationships, which are obtained
model simulation results, to convert streamflow attributes of
3-day averages back to daily ones. Values shown in parentheses
in Table 2 are daily flow statistics obtained through regression
relationships shown in Fig. 9. Similar regression relationships
could be obtained for other flow attributes and their statistical
properties. For instance, the FDC for the daily flow could be
obtained from the 3-day average FDC as follows:
1. Divide the FDC into segments by using the exceedance
probabilities as the breakpoints;
2. From the model simulations compute the flows \((Q)\) corre-
sponding to each exceedance probability \((p)\), \(Q_p\), for daily
and 3-day average flows. Repeat this for each MC simulation
result;
3. Regress \(Q_p\) values obtained from daily and 3-day average
flows for each break point \((p)\); and
4. Use the regression equations obtained in step (3) as a base to
convert the synthesized 3-day average flows to daily flows.
By following the above procedure we converted the ensemble
of 3-day average FDC shown in Fig. 7(a) to the ensemble of daily
FDC [Fig. 7(b)]. The exceedance probabilities of 0.001, 0.01,
0.05, 0.1, 0.2, . . . , 0.8, 0.9, 0.95, 0.99, 0.999 are used as the
breakpoints, and regression equations are generated at these loca-
tions with \(R^2=0.93\). Of course, the ideal procedure is to construct
such regression relationships from measured streamflows,
provided that adequately long historical records are available.
Unfortunately, this is not the case here. Comparison of the two

![Fig. 7. Median and 95% confidence band for MC simulated duration curve: (a) 3-day averaged flows; (b) daily flows. Flow duration curves are generated from 500 replicas of 20 years daily streamflows, \(\hat{o}_t\).](image)

![Fig. 8. Coefficient of variation computed from ensemble of flow duration curves versus probability of exceedance \(P(X \geq x)\).](image)
However, the often cited requirement of normally distributed white noise, \( w_t \), has precluded the use of times series models, especially in the class of hydrologic problems characterized by non-Gaussian random terms. This may be partly perpetrated by the inadequacy of forecasting/synthesis algorithms in readily available statistical packages to handle non-Gaussian \( w_t \). This limitation is relaxed in this paper by fitting a nonparametric distribution to the random component of the residual error model \([w_t \text{ in Eq. (3)}]\) and inverting the resulting CDF [Eq. (25)] to generate independent \( w_t \) values sampled from generic distributions. Excluding synthesized negative streamflow values, the computed 95% confidence bands, nonetheless, were large enough to encompass most of the observed values. The pitfall in the blind application of the \( l \)-step forecast without challenging the normality assumption can lead to inadequate estimation of uncertainty. Although, not strongly accentuated by this application, underestimation of uncertainty band can lead to a false sense of model reliability, as depicted by the narrower uncertainty band for the \( l \)-step forecast (Fig. 4). Figs. 5 and 6 show that for the semiparametric method most of the observed streamflows fell well within their respective 95% confidence intervals.

The higher COV values for low flows (Fig. 8) point toward significant uncertainties in the SWAT simulated low flows. This corroborates the higher COV values obtained by Kalin and Hantush (2006a) for the low-flow attributes, base flow, and \( 7Q10 \) relative to the COVs of median and high-flow metrics. The relatively high uncertainty was also observed for small rainfall events in a different model study in which the physically based runoff and sediment erosion model KINEROS yielded higher COV values during small rainfall events (Hantush and Kalin 2005). Since baseflow dominates streamflow during low-flow periods, the relatively high uncertainty in the simulated low flows might be attributed to inadequate calibration of the groundwater component of the SWAT model and/or errors arising from the baseflow separation method. Different methods of baseflow separation can give significantly different baseflow estimates; this clearly affects both calibrated model parameters and model performance. It should be noted that baseflow separation is the first step recommended in SWAT model calibration, and the less stringent \( \pm 15\% \) of baseflow volume error requirement recommended by Santhi et al. (2001) with no threshold values for the Nash–Sutcliffe efficiency and coefficient of determination highlight the dilemma in attempting to calibrate a quantity whose measurement itself is uncertain. Although not so evident, the proposed methodology may also be inadequate to handle low flows.

In general, the relatively low COV values (Table 2) indicate that the calibrated SWAT model is capable of predicting the streamflow attributes fairly reliably, especially when compared to predictions of real-time flow values. However, the relatively low COV of model forecasted annual maximum daily flow indicates a level of uncertainty to be communicated in terms of a greater risk factor for the performance flood control measures designed based on this metric, whereas the relatively low uncertainty associated with median monthly daily flow indicates that the calibrated model reliably estimates this ecologically significant indicator. The average daily flow, which also measures base flow during interstorm periods, showed an almost similar COV value to that for the average monthly median of daily flows.

**Fig. 9.** Conversion of 3-day averages to daily values of flow attributes

Because of their relative simplicity, classical linear time series models (e.g., Shumway 1988; and Box and Jenkins 1970) have been widely used in hydrology, economics, and many other fields. However, the often cited requirement of normally distributed white noise, \( w_t \), has precluded the use of times series models, especially in the class of hydrologic problems characterized by non-Gaussian random terms. This may be partly perpetrated by the inadequacy of forecasting/synthesis algorithms in readily available statistical packages to handle non-Gaussian \( w_t \). This limitation is relaxed in this paper by fitting a nonparametric distribution to the random component of the residual error model \([w_t \text{ in Eq. (3)}]\) and inverting the resulting CDF [Eq. (25)] to generate independent \( w_t \) values sampled from generic distributions. Excluding synthesized negative streamflow values, the computed 95% confidence bands, nonetheless, were large enough to encompass most of the observed values. The pitfall in the blind application of the \( l \)-step forecast without challenging the normality assumption can lead to inadequate estimation of uncertainty. Although, not strongly accentuated by this application, underestimation of uncertainty band can lead to a false sense of model reliability, as depicted by the narrower uncertainty band for the \( l \)-step forecast (Fig. 4). Figs. 5 and 6 show that for the semiparametric method most of the observed streamflows fell well within their respective 95% confidence intervals.

The higher COV values for low flows (Fig. 8) point toward significant uncertainties in the SWAT simulated low flows. This corroborates the higher COV values obtained by Kalin and Hantush (2006a) for the low-flow attributes, base flow, and \( 7Q10 \) relative to the COVs of median and high-flow metrics. The relatively high uncertainty was also observed for small rainfall events in a different model study in which the physically based runoff and sediment erosion model KINEROS yielded higher COV values during small rainfall events (Hantush and Kalin 2005). Since baseflow dominates streamflow during low-flow periods, the relatively high uncertainty in the simulated low flows might be attributed to inadequate calibration of the groundwater component of the SWAT model and/or errors arising from the baseflow separation method. Different methods of baseflow separation can give significantly different baseflow estimates; this clearly affects both calibrated model parameters and model performance. It should be noted that baseflow separation is the first step recommended in SWAT model calibration, and the less stringent \( \pm 15\% \) of baseflow volume error requirement recommended by Santhi et al. (2001) with no threshold values for the Nash–Sutcliffe efficiency and coefficient of determination highlight the dilemma in attempting to calibrate a quantity whose measurement itself is uncertain. Although not so evident, the proposed methodology may also be inadequate to handle low flows.

In general, the relatively low COV values (Table 2) indicate that the calibrated SWAT model is capable of predicting the streamflow attributes fairly reliably, especially when compared to predictions of real-time flow values. However, the relatively low COV of model forecasted annual maximum daily flow indicates a level of uncertainty to be communicated in terms of a greater risk factor for the performance flood control measures designed based on this metric, whereas the relatively low uncertainty associated with median monthly daily flow indicates that the calibrated model reliably estimates this ecologically significant indicator. The average daily flow, which also measures base flow during interstorm periods, showed an almost similar COV value to that for the average monthly median of daily flows.

**Conclusions**

A two-step methodology was presented for the estimation of predictive uncertainty of a distributed hydrologic model calibrated for the rapidly urbanizing Pocono Creek watershed. The first step requires adequate model calibration and the second, which is the topic of this paper, deals with stochastic residual-error analysis. The aim of the stochastic error analysis was the estimation of
uncertainty associated with daily streamflows and related long-term attributes simulated by the distributed hydrologic model. The methodology combined classical linear time series analysis (ARIMA) with a nonparametric probabilistic sampling methodology to synthesize an ensemble of serially correlated model residual errors of simulated streamflows by means of Monte Carlo-type Latin-hypercube simulations. The ensemble residual error time series was used to construct the forecast and long-term daily flow attributes of management and ecological significance, and flow duration curves for stream gauge station upstream from the mouth of the watershed. However, systematic prediction errors generated by the watershed model during significant events precluded the analysis using daily streamflow data. Alternatively, 3-day averages (as a surrogate to daily flow) were used to mitigate the effect of large errors during significant events and construct the time series model. Conversion of flow metrics from 3-day averages to daily was achieved by means of simple linear regression relationships.

An ARIMA time series model was fit to 3-day averaged model residual errors and a split-sample approach was implemented to construct and validate the error model. Two methods were used to construct the uncertainty band for model forecast, the 1-step forecast and the semiparametric model. The former relies on the assumption of Gaussian white noise (random term in the linear time series model), while the latter is independent of the distribution of the noise term and, therefore, is more versatile. It was shown that the application of the 1-step forecast could lead to an underestimation of the uncertainty band when compared to the wider band generated by the semiparametric model; however, both methods produced comparable results. Monte Carlo simulations showed that long-term annual maximum daily flow had the highest uncertainty (measured by COV) among four flow attributes. The long-term monthly median of daily flows (a wild trout habitat metric) and average daily flow (base flow during interstorm periods) showed lower uncertainty. The higher coefficient of variation (COV) values associated with simulated high-flow attributes indicate a level of uncertainty to be communicated in terms of a greater risk factor for the design and performance of flood control measures. The versatility of the synthesis methodology—by means of deterministic SWAT and stochastic residual error model simulations—was further highlighted by the construction of ensemble flow duration curves, which allowed for the estimation of the (range) of flow for a specified design recurrence (return) period with 95% confidence. The simulated low flows (<2 m³/s) at the site showed highest uncertainty, which may be attributed to a combination of both inadequate model account for groundwater discharge and errors in the basflow separation method, or inadequacy of the methodology for low flows. The computed streamflow prediction uncertainties and the associated flow attributes provide the basis for communicating the risk to water resources managers and decision makers to examine alternative management plans for sustainable water resources in the watershed.

The presented results, while specific to the study watershed, demonstrated a methodology that combines the relative simplicity of classical time series models with the versatility of nonparametric probabilistic analysis to account for all sources of model and data uncertainty. The approach, however, is suitable for continuous time simulations and contingent upon a successful model calibration.

Acknowledgments

The U.S. Environmental Protection Agency through its Office of Research and Development partially funded and managed the research described here through in-house efforts and in part by an appointment to the Postgraduate Research Program at the National Risk Management Research Laboratory. This program is administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. Environmental Protection Agency. It has not been subjected to Agency review and therefore does not necessarily reflect the views of the Agency, and no official endorsement should be inferred.

References


Hossain, F., and Anagnostou, E. N. (2005). “Assessment of a probabilis-