ISEIS

Journal of Environmental Informatics 19(2) 70-78 (2012)

Journal of Environmental Informatics

www.iseis.org/jei

Evaluation of Bayesian Estimation of a Hidden Continuous-Time Markov Chain Model with Application to Threshold Violation in Water-Quality Indicators

F. A. Deviney Jr[1,*], D. E. Brown[2], and K. C. Rice[3]

[1]*Department of Environmental Sciences, University of Virginia, Charlottesville, Virginia, USA*
[2]*Department of Systems and Information Engineering, University of Virginia, Charlottesville, Virginia, USA*
[3]*U. S. Geological Survey, Charlottesville, Virginia, USA*

Received 25 Feburary 2011; revised 21 Feburary 2012; accepted 9 April 2012; published online 22 June 2012

ABSTRACT. Natural resource managers require information concerning the frequency, duration, and long-term probability of occurrence of water-quality indicator (WQI) violations of defined thresholds. The timing of these threshold crossings often is hidden from the observer, who is restricted to relatively infrequent observations. Here, a model for the hidden process is linked with a model for the observations, and the parameters describing duration, return period, and long-term probability of occurrence are estimated using Bayesian methods. A simulation experiment is performed to evaluate the approach under scenarios based on the equivalent of a total monitoring period of 5-30 years and an observation frequency of 1-50 observations per year. Given constant threshold crossing rate, accuracy and precision of parameter estimates increased with longer total monitoring period and more-frequent observations. Given fixed monitoring period and observation frequency, accuracy and precision of parameter estimates increased with longer times between threshold crossings. For most cases where the long-term probability of being in violation is greater than 0.10, it was determined that at least 600 observations are needed to achieve precise estimates. An application of the approach is presented using 22 years of quasi-weekly observations of acid-neutralizing capacity from Deep Run, a stream in Shenandoah National Park, Virginia. The time series also was sub-sampled to simulate monthly and semi-monthly sampling protocols. Estimates of the long-term probability of violation were unbiased despite sampling frequency; however, the expected duration and return period were over-estimated using the sub-sampled time series with respect to the full quasi-weekly time series.

Keywords: Bayesian MCMC methods, CTMC, Shenandoah National Park, threshold violation, water quality, hidden processes

1. Introduction

Aquatic resource professionals recognize that harm can come to aquatic ecosystems when water-quality thresholds are violated too often or for excessively long periods of time (Davies et al., 1992; DeWalle et al., 1995; Sickle et al., 1996; Baldigo and Murdoch, 1997; Bulger et al., 2000; Laio et al., 2001). These thresholds may be set by regulation or be determined empirically by studies. Either side of the threshold may represent the harmful condition.

For water quantity, i.e., river discharge, hydrologists have worked for decades to couple extreme value theory with the theory of stochastic processes. The motivation has been the need to be able to predict the likelihood of occurrence of infrequent and random flood events, within typically long management horizons such as the lifetime of a structure (e.g., a bridge). For these analyses, relatively high-frequency (with respect to period), evenly spaced observations of discharge (e.g., 15-minute or hourly observations) are available. These high-frequency observations allow the timing of threshold crossings to be identified with little uncertainty relative to the occurrence of crossings of interest.

While relatively high-frequency observations are available for discharge, this is not the case for water-quality indicators (WQI), where the frequency of observation is typically monthly, quarterly, or rarely, weekly. This is especially true for long-term monitoring programs. In these cases the exact time, or even the approximate time, of a threshold crossing cannot be known with any reasonable degree of certainty. Likewise, the time between crossings cannot be known with reasonable certainty.

Since these crossings generally are not observed, the process that generates them is a hidden process. Hidden processes are common in the literature and are sometimes referred to as latent or state processes, and the variables associated with them are referred to as hidden, latent, or state variables (Berliner, 1996; Skrondal and Rabe-Hesketh, 2004; Cressie and Wikle, 2011). The key with hidden processes is to identify a mathematical mapping between what can be observed and the true process and/or variables of interest.

Thresholds of interest differ between analyses of water quantity and water quality. In flood frequency analysis, the th-

* Corresponding author. Tel.: +1 434 9717683. *E-mail address:* deviney@virginia.edu (F. A. Deviney Jr).

ISSN: 1726-2135 print/1684-8799 online
 doi:10.3808/jei.201200210

resholds are usually relative. For example, in the Peaks Over Threshold (POT) method (Todorovic and Zelenhasic, 1970), the threshold for determining when crossings occur typically is set as a high percentile value. For WQIs, however, thresholds of interest usually are specified as absolutes. WQI thresholds are not necessarily near the top of the range of observations, as flood-frequency thresholds are; in fact, WQI thresholds might be near the bottom of the range of observations. For example, in this paper we examine a WQI threshold defined as “acid-neutralizing capacity” (ANC) equal to zero, because when ANC is negative (below the threshold), the water has lost its ability to neutralize acidity and thus can be harmful to aquatic organisms (Bulger et al., 2000). Periods of ANC depression less than zero of only a few days can be critical, yet ANC is often only observed on a weekly or monthly time step. Under these conditions, the observer will rarely know when threshold crossings occur.

The new contributions relative to the modeling of threshold violations in WQIs presented here are threefold. First, we related a model for observations of a WQI to a model for the threshold-crossing process, which is hidden from the observer but is what is of interest, and developed a procedure for estimating those parameters using Bayesian methods. This model and method will be of value to water-quality analysts interested in the frequency, duration, and long-term probability of occurrence of threshold-violation events. Second, we conducted a simulation experiment to determine the relationship between accurate and precise parameter estimation and monitoring frequency and period. The understanding resulting from this experiment will be of value to those conducting monitoring. Third, we applied the methodology to a time series of data collected from a stream in Shenandoah National Park to examine the utility of the method in a real setting. In the remainder of this paper we introduce the modeling and estimation machinery behind the method, describe the experiment conducted to validate its use, and analyze the results of the experiment. We then provide the example application, followed by discussion of the limitations of the method, opportunities for additional work, and recommendations for practitioners. We conclude the paper with a summary of our contributions.

2. Method

Markov processes have been widely discussed in hydrology (Lu and Berliner, 1999; Szilagyi et al., 2006). A number of researchers have developed models for threshold violations that combine extreme value and Markov chain theory (Smith et al., 1997). Others have used statistical (Deviney et al., 2006) or process (Zhang and Arhonditsis, 2008) models to make time series predictions from which threshold violation properties can be estimated. However, these approaches assume the existence of a high-frequency time series of either the WQI or of another variable, i.e., discharge, which can be used to make high-frequency predictions of the WQI.

A homogeneous two-state continuous-time Markov chain (CTMC) is a reasonable first choice for a model of threshold violation, where the times of threshold crossings are hidden from the observer. To describe this model, let state 'A' represent the violation state and let state 'B' represent the non-violation state. Assume that upon entering state 'A', a random amount of time X, the duration, passes before the process transits to state 'B'. Once in state 'B', a random amount of time Y, independent of X, passes before the process returns to state 'A'. Assume these random times have exponential distributions with rates λ and μ that are not necessarily equal. Such a process has the following well-known properties:

$$\begin{aligned} & f(x;\lambda) = \lambda e^{-\lambda x}, x \geq 0 \\ & f(y;\mu) = \mu e^{-\mu y}, y \geq 0 \\ & E[X] = \frac{1}{\lambda} \\ & E[Y] = \frac{1}{\mu} \\ & E[Z] = E[X] + E[Y] \\ & P_{\mathrm{A}} = \frac{\mu}{\lambda+\mu} \\ & P_{\mathrm{B}} = \frac{\lambda}{\lambda+\mu} \end{aligned} \tag{1}$$

where f($x;\lambda$) and f($y;\mu$) are the probability density functions of X and Y, X represents the random amount of time spent in state 'A', Y represents the random amount of time spent in state 'B', Z = X + Y is the random time between entrances into either state (the return or renewal period), E[·] denotes expectation, P_A is the long-term probability of occurrence of state 'A' (alternatively the long-term proportion of time spent in state 'A' or the limiting probability for state 'A'), and P_B is the long-term proportion of time spent in state 'B'. If one knew when state transitions (threshold crossings) occurred, it would be an easy matter to make estimates of these properties, which are the properties of interest to the WQI analyst.

Ross (2006) showed how the two-state homogeneous CTMC process and the Kolmogorov Backward Equations lead to two equations that predict the probability of being in state “A” 1) having been in state “A”, or 2) having been in state “B”, at some time previous. The Kolmogorov Backward Equations are differential equations that relate transition rates and transition probabilities between states. In the case of the two-state homogeneous CTMC, Ross (2006) showed how these differential equations can be solved rather simply to yield the following:

$$\begin{aligned} & P_{\mathrm{AA}}(\Delta t) = \frac{\lambda}{\lambda+\mu} e^{-(\lambda+\mu)\Delta t} + \frac{\mu}{\lambda+\mu} \\ & P_{\mathrm{BA}}(\Delta t) = \frac{\mu}{\lambda+\mu} - \frac{\mu}{\lambda+\mu} e^{-(\lambda+\mu)\Delta t} \end{aligned} \tag{2}$$

where P_{AA} (Δt), for example, is the probability of going from state 'A' to state 'A' in Δt time, over all possible transitions through state 'B' in the meantime. The elegance of these equations can be seen by considering the cases where Δt goes to zero or to infinity. In the first case, $P_{AA} = 1$ and $P_{BA} = 0$, which is to be expected given that not much time has elapsed. In the second case, $P_{AA} = P_{BA} = \mu/(\lambda+\mu)$, which is just the long-term probability of being in state 'A'. In other words, the influence of the last observed state on the current state diminishes with time until it is no longer felt. Similar results can be derived for P_{BB} and P_{AB}.

While the analyst is interested in the distributions of times between threshold crossings, which are governed by the parameters λ and μ, observations are generally only available between crossings. Equations (1) and (2) link the process parameters with the observations. In addition, they impose no requirement that the observations be evenly spaced, frequent, or coincident with state changes. With some re-arrangement of terms, a hierarchical specification for the model can be written as:

$$
\begin{aligned}
&Prob\left(S_n = \text{'A'}\right) \sim Bernoulli\left(P_{\text{A}} F_n + \left(1 - F_n\right) Prob\left(S_{n-1} = \text{'}A\text{'}\right)\right) \\
&Prob\left(S_{n-1} = \text{'}A\text{'}\right) = \begin{cases} 1, & S_{n-1} = \text{'}A\text{'} \\ 0, & S_{n-1} = \text{'}B\text{'} \end{cases} \\
&P_{\text{A}} = \lambda / \left(\lambda + \mu\right) \\
&F_n = 1 - e^{-(\lambda+\mu)\Delta t_n} \\
&\Delta t_n = t_n - t_{n-1}
\end{aligned}
\qquad (3)
$$

where λ and μ are the transition rates from state 'A' to 'B' and 'B' to 'A', respectively, and S_n is the state at time index n.

It is common to take WQI measurements at semi-regular intervals, either in situ with a meter or by collecting a sample for laboratory analysis. While in actual practice the value of the WQI is recorded, for analysis with respect to a biological or regulatory threshold, for this model it is only necessary to know whether or not the value was above or below the threshold. We wanted to simulate this semi-regular observation behavior. A method was developed to simulate the repeated measurement, or observation, of a WQI, where the observation is simply whether or not the WQI is above or below a defined threshold value. Each simulation was then used as the input to the estimation procedure for the parameters λ and μ of the two-state homogeneous CTMC described in equations (1), (2), and (3).

For the estimation procedure we chose a hierarchical Bayes framework. While not employed in this case beyond simple observation and process equations, the hierarchical Bayes framework allows the process parameters to be predictted by exogenous variables, such as watershed characteristics. Such extensions lead to a potentially infinite variety of non-homogeneous CTMC models. The intention here was to establish first that given a simple process model, the estimation procedure would yield the true (design) values of the process parameters, and second, establish the rates and numbers of observations needed to yield the true (design) values with precision. Without such an analysis, practitioners would have little reason to trust the results of using the method with real data or with more complicated extensions to the model.

An experimental design was created for testing the estimation procedure. Twenty-five hundred sets of design parameter values were generated. For each parameter set, two separate simulations were generated. In total, 5,000 sets of simulated observations were created. Simulations of a WQI were generated using the CTMC process model described above in equation (2). Simulations were designed to exhibit a range of WQI processes typical of, and of interest in, the water-quality monitoring field (Table 1). State 'A' in the previous model descriptions was arbitrarily chosen to be the violation state. Parameters were set so that the design mean violation state duration (*E[X]*) for each simulation was chosen from an interval between 0.005 periods and 0.05 periods (given a period of one year, this range is roughly between 40 hours and 2.5 weeks). Crossings into the violation state in each simulation had a design mean return period (*E[Z]*) chosen from between 1 and 0.1 (between once per period and ten times per period, or roughly from annually to monthly). These ranges for *E[X]* and *E[Z]* result in a range for PA of between 0.005 and 0.5.

Table 1. Ranges of Important Process Characteristics of CTMCs Used in Simulations

Property	Range	Description
E[X]	[0.005, 0.05]	Expected length of time in the violation state
E[Z]	[0.1, 1.0]	Expected time between entries into the violation state, or the mean return period

Table 2. Ranges of Important Process Observation Properties

Property	Range	Description
oP	[5, 50]	Observation period length
rS	[0.01, 1.0]	Quasi-regular observation interval (spacing is equal ± a small perturbation)

WQI observations were simulated at a spacing of 0.0001 time units apart (10,000 data points per period). Relative to the solar year, that is roughly equivalent to 1 data point per hour. It was assumed that more frequently simulated observations would not be necessary to capture state changes important in water-quality monitoring.

Observation of the WQI process was simulated by quasi-randomly selecting observations from those described above, based on an *observation protocol*. Two minor components were defined associated with the observation protocol (Table 2). For the first component *observation period length, oP*, lengths of between 5 and 50 periods were specified, as the lower limit corresponds roughly to a generally accepted minimum time period for analysis of stochastic processes with a seasonal component (Hirsch et al., 1982), and the upper limit exceeds but is within the foreseeable range of observed data records available in water-quality monitoring. The second minor component of the observation protocol *quasi-regular observation interval, rS,* mimicked a regular collection spaced at intervals of 0.01 to 1 period (100 to 1 observations per period), where the actual interval lengths were allowed to vary with a standard deviation equal to one fifth of the mean interval length. For example, for a mean interval of one week (168 hours), the standard deviation would have been set to 33.6 hours.

Choices for design parameter levels describing the processes and observation protocols were generated using leaped Halton sequences (Kocis and Whiten, 1997). Halton sequences are used in computer experiments, where replication is usually not possible, to obtain good coverage over the design space. To ensure orthogonality between design parameters, the leaped Halton sequence for each parameter was generated using a unique prime number as the base, and a common leap parameter of a prime number greater than the maximum base value.

For each of the 5,000 sets of simulated observations, posterior distributions of model parameters were estimated using OpenBUGS (Thomas et al., 2006), a Bayesian estimation software package that can be run from R using the BRugs package (Thomas et al., 2006). All estimates were made using the Cross-Campus Grid (XCG) at the University of Virginia (Morgan and Grimshaw, 2007). Convergence of each set of estimates was determined by requiring that the Gelman-Rubin statistic $\hat{R}$ be less than 1.01 for each monitored parameter. Gelman and Hill (2007) suggest a threshold of 1.1 for preliminary work, and smaller values for more stable estimates. In addition, the slopes of all chains were required to be trend free. To check this, a regression of chain values versus their iteration index, modeling residuals to be first-order autocorrelated, was performed. A conservative p-value of 0.01 divided by the total number of chains was used for rejection of the null hypothesis of "no trend" to compensate for the number of tests.

Estimates of each set of posterior distributions were made using three chains. After burn-in, sufficient iterations were run to obtain 350 values from each chain, for a total of 1,050 samples for the posterior distributions of λ and μ. Each estimation run was initialized with a relatively small burn-in period and thinning parameter. Then, if convergence was not attained, the thinning parameter was increased and the procedure ran again, starting from the last iteration's values of the previous run, and considering all iterations of the algorithm up to that point to be burn-in.

The Gamma distribution (equation (4)) is a natural conjugate prior for the exponential distribution rate parameters λ and μ (Gelman et al., 2004):

$$\lambda, \mu \sim Gamma(\alpha, \beta)$$
$$p(x) \sim \frac{\beta^{\alpha} x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \tag{4}$$

The hyper-parameters for the Gamma priors were chosen (both α and β were set to 0.001) to provide relatively flat, but weakly informative, priors. The prior expectation and variance for λ and μ were thus 1 and 1,000, respectively, making the prior expectations for *E[X]*, *E[Z]*, and P_A, respectively, 1, 2, and 0.5. These are outside of, or at the edge of, the experimental ranges.

The three design parameters assessed for recovery and precision were: 1) *E[X]*, the expected value of time spent in the violation state, 2) *E[Z]*, the expected value of the renewal interval (time between re-entries to the violation state, or the inverse of frequency of occurrence), and 3) P_B, the limiting probability for the non-violation state. Recovery was evaluated by examining whether the design parameter value was within a 95% credible interval determined from the posterior distribution for the parameter. A 95% credible interval for the posterior was defined as the values corresponding to the percentiles between 2.5 and 97.5.

Recovery can be obtained with a wide credible interval, however, a narrow credible interval generally is desired. Precision was examined with a *relative dispersion* (*RD*) metric, calculated as:

$$RD = \frac{95\%\ credible\ interval\ width}{4x\ estimated\ \ parameter\ value} \tag{5}$$

where the estimated parameter was taken as the median of the posterior distribution. *RD* is approximately equivalent to the relative standard deviation (*RSD*) if one assumes that a 95% credible interval is approximately four standard deviations in width.

3. Results

All of the 5,000 sets of estimates converged successfully. Close to 98% of the design parameter percentiles fell within their respective credible intervals (Table 3).

Table 3. Counts of Design Values with Respect to a 95% Credible Interval of the Posterior Distribution

Group	# inside c.i.	total	% inside c.i.
E[X]	4,943	5,000	98.86
E[Z]	4,941	5,000	98.82
P1	4,840	5,000	96.80
All	14,724	15,000	98.16

The relative dispersion, *RD,* was calculated for estimates of *E[X], E[Z]*, and P_B. In order to gain some intuition about any relationship between *RD* and the four factors being manipulated in the experiment, the *RD* metrics (log-transformed) were regressed against suitable transformations of the four factors (Table 4). Although in each case there is some unexplained variation in the metric, the models are statistically significant. Inclusion of second-order effects did not improve the models sufficiently to warrant the increased model complexity. Of interest is the not unexpected result that in all three cases, *RD* improves (decreases) with increasing observation period length or shorter observation intervals. Of note also is that the results for *E[X]* and *E[Z]* are virtually identical (Table 3).

These results can be visualized by holding *E[X], E[Z]*, and *RD* constant while allowing *oP* and *rS* to vary. Figure 1 illustrates the relationship between *rS* and *oP* for various given values of $RD_{E[X]}$, *E[X],* and *E[Z]* in each subplot. These plots indicate that fewer observations are required to obtain small

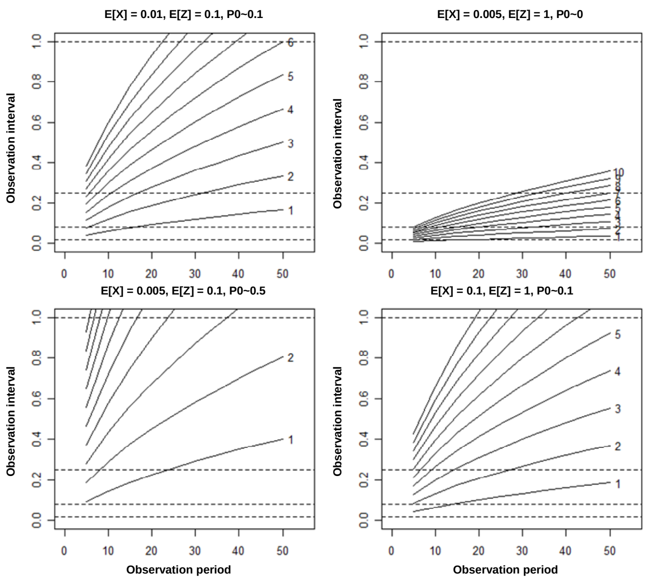

Figure 1. Contours of $RD_{E[X]}$ are plotted for observation interval vs. observation period for various combinations of E[X] and E[Z]. Dashed horizontal lines correspond to weekly, monthly, and quarterly observation intervals, when the period is one year.

Table 4. Regression Results

RD	Regression results				
$RD_{E[X]}$	Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
	(Intercept)	2.92121	0.08637	33.82	< 2e-16 ***
	log_E_X	-0.54793	0.01717	-31.92	< 2e-16 ***
	log_E_Z	0.50394	0.01717	29.35	< 2e-16 ***
	log_oP	-0.63710	0.01716	-37.12	< 2e-16 ***
	log_rS	1.00234	0.01140	87.95	< 2e-16 ***
	Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				
	Residual standard error: 0.7136 on 4995 degrees of freedom				
	Multiple R-squared: 0.6875, Adjusted R-squared: 0.6873				
	F-statistic: 2747 on 4 and 4995 DF, p-value: < 2.2e-16				
$RD_{E[Z]}$	Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
	(Intercept)	2.79089	0.08448	33.03	< 2e-16 ***
	log_E_X	-0.52869	0.01679	-31.48	< 2e-16 ***
	log_E_Z	0.48444	0.01679	28.84	< 2e-16 ***
	log_oP	-0.59740	0.01679	-35.58	< 2e-16 ***
	log_rS	0.98202	0.01115	88.09	< 2e-16 ***
	Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				
	Residual standard error: 0.698 on 4995 degrees of freedom				
	Multiple R-squared: 0.6847, Adjusted R-squared: 0.6844				
	F-statistic: 2711 on 4 and 4995 DF, p-value: < 2.2e-16				
RD_{P1}	Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
	(Intercept)	0.334363	0.047452	7.046	2.09e-12 ***
	log_E_X	0.264793	0.009432	28.073	< 2e-16 ***
	log_E_Z	-0.334707	0.009433	-35.482	< 2e-16 ***
	log_oP	-0.737217	0.009431	-78.173	< 2e-16 ***
	log_rS	0.558216	0.006262	89.147	< 2e-16 ***
	Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				
	Residual standard error: 0.392 on 4995 degrees of freedom				
	Multiple R-squared: 0.7635, Adjusted R-squared: 0.7633				
	F-statistic: 4031 on 4 and 4995 DF, p-value: < 2.2e-16				

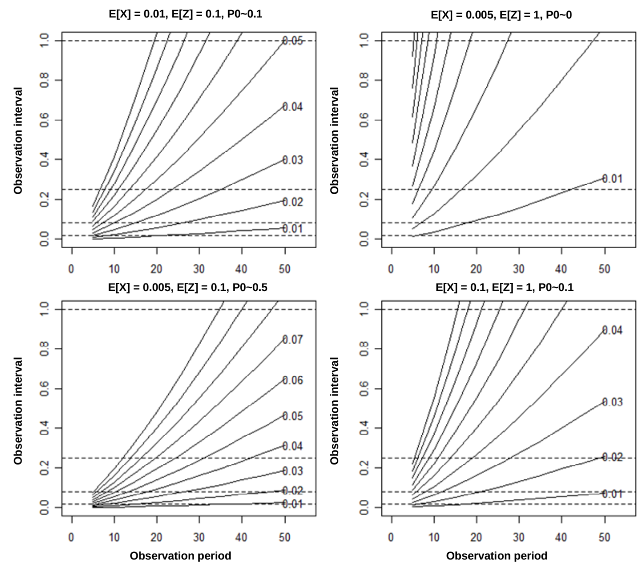

Figure 2. Contours of RD_{PB} are plotted for observation interval vs. observation period for various combinations of E[X] and E[Z].

RD when the proportions of time spent above or below the threshold are nearly equal (lower left subplot) than when these proportions are unequal. The upper left and lower right subplots suggest that given two processes with equal P_A, it will require more observation effort to get the same *RD* for the process with shorter return period (higher transition rates). For P_A = 0.10, a common proportion for legal impairment designation of a water body, it appears that decades of observations at weekly or monthly intervals would be necessary to obtain an *RD* much less than 1. The plot for $RD_{E[Z]}$ is similar to Figure 1.

In contrast, Figure 2 indicates that for P_B (and by inference P_A), *RD* is substantially smaller than for either *E[X]* or *E[Z]* given the same observation effort. In other words, the proportions will always be known with greater certainty than the mean duration or the mean return period. This is not unexpected since there can be many pairs of values of *E[X]* and *E[Z]* that yield the same proportion values.

In both Figure 1 and Figure 2, lines that pass through the origin represent scenarios of equal numbers of observations, because the number of observations is equal to the observation period length divided by the observation interval length. To achieve *RDs* for *E[X]* and *E[Z]* that are less than 1 when $P_A \sim 0.5$, 0.1, or < 0.01, about 150, 300 or 1,000 observations, respectively, are needed. These results are based on an assumption that the process is stationary over the period of record.

When the number of observations was small, design percentile values for small values of *E[X]* (less than about 0.03) were biased (not centered at the 50th percentile), even if they were within the 2.5 to 97.5 percentile range. When the *RD* calculated for the estimate was less than 2, however, no such bias was evident (Figure 3).

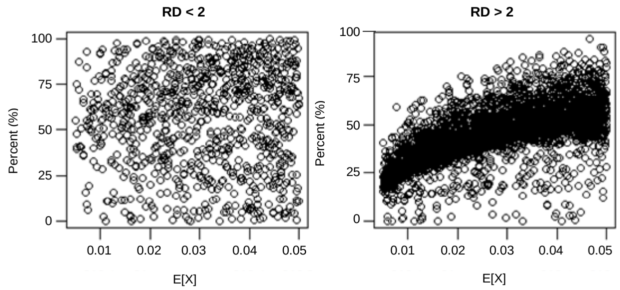

Figure 3. Distributions of nominal values of E[X] as percentiles of the posterior distributions when RD was < 2 (left subplot) and > 2 (right subplot).

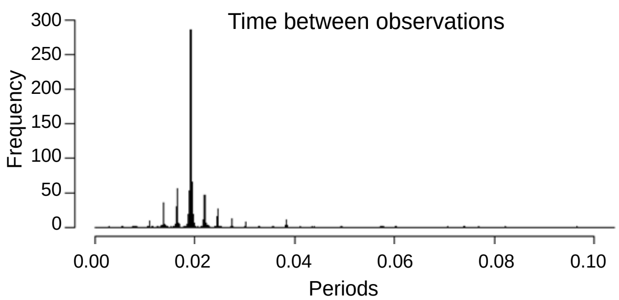

Figure 4. The spike at 0.02 periods is roughly one week (7 days). Spikes are observed at 4-11 days and at 14 days. Blips to the right of 0.04 indicate droughts or heavy snow periods when the station was inaccessible.

4. Application of the Model

Waters with low ANC are known to be harmful to aquatic life, and excursions of ANC of sufficient duration below zero microequivalents per liter (μeq/L) are thought to be sub-lethal (stressful) or lethal for some species (Bulger et al., 2000). A 26-year time series of 1,220 quasi-weekly observations of ANC was obtained for Deep Run, a small headwater stream draining a 306 hectare forested catchment in the southwest part of Shenandoah National Park, Virginia (Ryan et al., 1989). We used this dataset to examine the effect of different sampling frequencies on the precision and accuracy of parameter estimates and compared the results with our conclusions from the simulation experiment.

Deep Run is visited on a quasi-weekly basis, which means that it is visited at time intervals within a few hours of exactly 7 days apart. This interval, however, may be a few days more or less than seven days, or up to several months during drought or other severe weather conditions (Figure 4). The standard deviation of all observation interval lengths was ~ 0.004 periods (years) or ~ 1.5 days (statistics were restricted to intervals < 0.05 periods). The median of these restricted-interval data was 0.019 periods (7.0 days). The ratio of the standard deviation to the median interval was ~ 0.215. This ratio is nearly equivalent to that used to model observation interval length variability in the simulation experiment.

We converted the ANC time series to a binary series, where the value 1 indicates ANC ≥ 0 μeq/L and the value 0 indicates ANC < 0 μeq/L, which signifies the violation state. We used the entire dataset to estimate parameters using the method and model developed in section 2. We then simulated a quasi-monthly sampling protocol by dividing the year into 48 equal-length segments and assigning the observations from every fourth segment to a separate time series. This resulted in four time series. We analyzed each time series individually and ensembled the results by combining one-fourth of each of the posterior sample sets into a single posterior sample set. Finally, we simulated a semi-monthly sampling protocol by assigning the observations from every other segment to a separate time series. Each time series was analyzed and the posterior distributions combined as before.

In the simulation experiment, when estimate uncertainty was high (high values of *RD*), the mean duration period tended to be over-estimated. We suspected this bias arose from the choice of values for the hyper-parameters of λ and μ, which were intentionally set to be outside the experimental range. The fewer the number of observations, the closer the estimates will lie to the prior values.

In actual practice one could take advantage of the data to prescribe values for the hyper-parameters of the prior distributions of λ and μ. By counting the number of threshold crossings, one can make an estimate of *E[Z]*. Additionally, a rough estimate of $P_{violation}$ can be made from the proportion of observations in the violation state. From these two values, estimates of α and β for both λ and μ can be calculated (Equation (6)). For the analyses of Deep Run data, we prescribed the values of the hyper-parameters in this way:

$$
\begin{aligned}
&\hat{E}[Z] \approx \frac{oP}{(\#\,of\ threshold\ crossings)/2} \\
&\hat{P}_{\mathrm{A}} \approx \frac{\#\,of\ observations\ in\ violation}{total\ \#\,of\ observations} \\
&\hat{P}_{\mathrm{B}} = 1 - \hat{P}_{\mathrm{A}} \\
&\beta_{\lambda} = 1/\left(1000\hat{P}_{\mathrm{A}} E\left[\hat{Z}\right]\right) \\
&\alpha_{\lambda} = 1000\beta_{\lambda}^{2} \\
&\beta_{\mu} = 1/\left(1000\hat{P}_{\mathrm{B}} E\left[\hat{Z}\right]\right) \\
&\alpha_{\mu} = 1000\beta_{\mu}^{2}
\end{aligned}
\tag{6}
$$

The estimation processes converged within a few thousand iterations of the MCMC algorithm. The $\hat{R}$ values for both λ and μ were consistently less than 1.01, indicating good convergence. The parameter estimates using the entire dataset were quite precise, with *RDs* < 0.1. While this dataset had ~ 1,200 observations, the two quasi-semi-monthly datasets had ~ 600 each, and the four quasi-monthly datasets had ~ 300 observations each. The effect of these sampling frequencies on bias and precision can be seen clearly in Figure 5, in which cumulative distribution functions (CDFs) of the combined posterior distributions are plotted for *E[X]*, *E[Z]*, $P_{ANC<0}$, *and* $P_{ANC\geq 0}$. Exact values for medians, C.I.s, and *RD*s are given in Table 5. Precise estimates result in steep CDFs, such as is seen for the

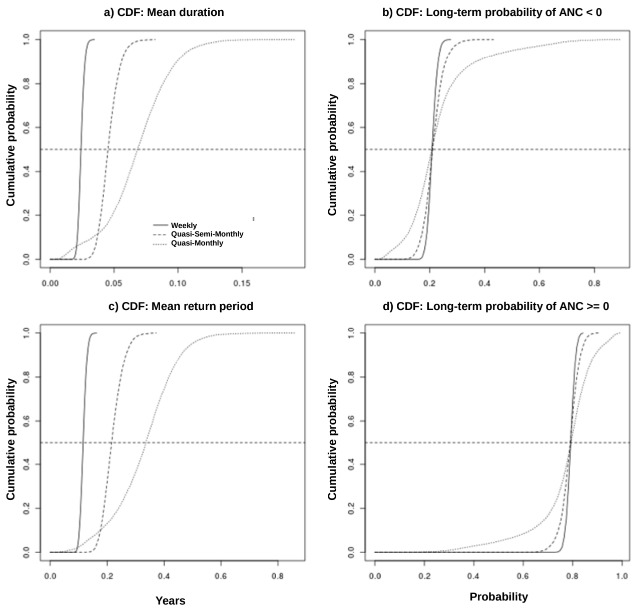

Figure 5. CDFs for properties of interest given quasi-weekly, quasi-semi-monthly, and quasi-monthly sub-sampling protocols.

Table 5. Summary of Results of Analysis of Deep Run ANC. E[X] and E[Z] Given in Years

	Parameter	Median	Credible interval	RD
Weekly	E[X]	0.024	[0.020, 0.029]	0.094
	E[Z]	0.116	[0.096, 0.139]	0.093
	$P_{ANC<0}$	0.208	[0.177, 0.242]	0.078
	$P_{ANC\geq 0}$	0.792	[0.758, 0.823]	0.021
Quasi-Semi-Monthly	E[X]	0.045	[0.033, 0.062]	0.167
	E[Z]	0.216	[0.160, 0.294]	0.155
	$P_{ANC<0}$	0.211	[0.145, 0.295]	0.178
	$P_{ANC\geq 0}$	0.789	[0.705, 0.855]	0.048
Quasi-Monthly	E[X]	0.068	[0.013, 0.118]	0.386
	E[Z]	0.336	[0.101, 0.531]	0.320
	$P_{ANC<0}$	0.207	[0.047, 0.618]	0.690
	$P_{ANC\geq 0}$	0.793	[0.382, 0.953]	0.180

posterior estimates using the entire dataset. While the true values of these quantities cannot be known, the results of the computer experiment suggest that the method is unbiased given sufficient observations. If we assume that the medians of the distributions of *E[X]* and *E[Z]* obtained using the entire dataset are sufficiently close to the truth, then the medians of the other distributions are clearly biased high in comparison. This agrees with the results found in the simulation experiment. In addition, the credible intervals for the quasi-monthly estimates are so wide as to be practically useless. For example, the credible interval for *E[X]* using the quasi-monthly datasets, in days, spans from roughly 4.7 to 43, a hardly useful measure in the context of exposure of biota to hazardous conditions. The *RD* for this credible interval was 0.386. In contrast, the credible interval for *E[X]* from the quasi-weekly data spanned, in days, from roughly 7.3 to 10.6, with an *RD* of 0.094.

The medians for the estimates of $P_{ANC<0}$, and $P_{ANC\geq 0}$ are roughly equal despite the sampling interval, although the uncertainty in those estimates increases with reduced sampling frequency. These results also agree with our conclusions from the simulation experiment.

5. Discussion

We are not aware of the development of any alternative methods for estimating the parameters of the model proposed in this study, when threshold crossings are hidden from the observer, so we have not made any comparisons. We expect to be able to extend the model to estimate parameters for multiple sites simultaneously, and to be able to model these parameters in terms of exogenous variables such as, for example, catchment area or geology, in a hierarchical modeling context. However, examining the properties of the estimation method itself in such a context would be difficult at best. We preferred to start out by examining the simple case of a single process.

The simulation experiment results indicated that between 150 and 300 observations would be needed to obtain relatively precise (relative precision less than 1) estimates of mean duration and return period. However, the Deep Run analysis results indicated that a relative precision, in terms of *RD*, on the order of 0.10 would be required to get estimates precise enough for management decisions. Figure 1 suggests that a minimum of around 600 observations (the equivalent of 50 years of monthly observations) would be needed to achieve this level of precision, although more than this amount was needed with the Deep Run data.

In comparison with trend analysis, more observations are needed to obtain satisfactory results. For example, Hirsch and Slack (1984) found that to obtain adequate power to detect trends, approximately 10 years of monthly observations, or 120 observations, were sufficient. Here we found that a bare minimum of 150 observations would be needed.

An important limitation of the model and method presented here is the assumption of process homogeneity. To be of greater use, the model needs to be modified to account for non-homogeneous processes, including seasonally varying behavior and changes over time. We expect to be able to develop multi-level models to incorporate extraneous explanatory variables, which should allow us to model the process parameters λ and μ as functions of time, season, or other watershed characteristics.

Another limitation of the model concerns the assumption that observations are made independently of the state of the process. Many water-quality monitoring projects include a mixture of occasional high-frequency observations made during high-discharge events along with regularly spaced observations. In these cases observations are generally not made independently of the state of the process. Deviney (2009) found that application of this model to simulations of such mixed observation protocols led to biased results. Observations made in conjunction with high-frequency observations need to be excluded from analyses using this methodology.

While five periods (if periods are years) is short for a typical long-term monitoring project, it is long for a typical research project. Therefore this method may not extrapolate well to shorter research projects, even given sufficient observations. There also may be situations where an event occurring much less frequently than once a period could have a significant impact on some ecosystems. While expected duration and return period were specified to be less than one, these results could be adapted to both such situations by a suitable definition of the length of one period. For example, if the unit period length was defined to be one week instead of one year, the implycations of Figure 1 would be the same adjusted to the new period unit.

6. Conclusions

Excursions of WQIs above or below certain thresholds for lengthy or frequent periods are known to have undesirable ecological consequences. However, estimation of parameters describing the return period, duration, and long-term probability of violation associated with such excursions is problematic because observations of WQIs are generally made infrequently, at arbitrary times, at unevenly spaced intervals, and asynchronously with threshold crossings.

Two-state continuous time Markov chains are simple processes that are useful for modeling a wide variety of phenomena. In this paper, we investigated the use of such models for estimating the parameters discussed above. A two-state continuous-time Markov chain (CTMC) model based on the Kolmogorov Backward Equations was evaluated using simulations of infrequent, unevenly spaced, and uncoordinated observations. Parameters were estimated using Bayesian methods. The evaluation consisted of an assessment of design parameter recovery and estimate precision, and of the effect of various process and observation characteristics on said precision. The purpose of these assessments was to determine the conditions under which the methodology could be applied, under the assumption that successful application would be a function of both process characteristics and of observation protocol characteristics.

Reasonably precise estimates of mean duration, return period, and long-term exceedance probability were obtained given sufficient observations. For most cases where the exceedance probability exceeds 0.10, at least 600 observations were determined to be needed. It is expected that many current water-quality time series meet this requirement. An example with over 1,200 quasi-weekly observations was presented. Very precise estimates of mean duration, return period, and long-term probability of violation were obtained with the full dataset. Precision deteriorated when the times series was sub-sampled at quasi-semi-monthly and quasi-monthly intervals. Bias was evident for estimates of mean duration and mean return period between the various rates of sampling, but was absent for estimates of the long-term probability of violation.

Additional research should test further extensions of these models, including non-homogeneous models and multi-level models, for single and multiple processes. Such models could take account of seasonal or trend changes in process parameters, make estimations for multiple locations simultaneously, or leverage other available information.

Acknowledgments. We thank Jim Tesoriero and Brent Troutman of the U.S. Geological Survey for their useful comments on an earlier draft of this paper. In addition, we acknowledge Shenandoah National Park and the Shenandoah Watershed Study at the University of Virginia for collecting and providing the data used in the example application.

References

Baldigo, B.P., and Murdoch, P.S. (1997). Effect of stream acidification and inorganic aluminum on mortality of brook trout (Salvelinus fontinalis) in the Catskill Mountains, New York. *Can. J. Fish. Aquat. Sci.,* 54(3), 603-615. http://dx.doi.org/10.1139/f96-314

Berliner, L.M. (1996). Hierarchical Bayesian time-series models. *Fundam. Theor. Phys.,* 79, 15-22.

Bulger, A.J., Cosby, B.J., and Webb, J.R. (2000). Current, reconstructed past, and projected future status of brook trout (Salvelinus fontinalis) streams in Virginia, *Can. J. Fish. Aquat. Sci.,* 57(7), 1515-1523. http://dx.doi.org/10.1139/f00-086

Cressie, N.A.C., and Wikle, C.K. (2011). Statistics for spatio-temporal data, Wiley.

Davies, T.D., Tranter, M., Wigington, P.J.Jr, and Eshleman, K.N. (1992). "Acidic episodes" in surface waters in Europe. *J. Hydrol. (Amst.),* 132(1-4), 25-69. http://dx.doi.org/10.1016/0022-1694(92)90172-R

Deviney, F.A., Rice, K.C., and Hornberger, G.M. (2006). Time series and recurrence interval models to predict the vulnerability of streams to episodic acidification in Shenandoah National Park, Virginia, *Water Resour. Res.,* 42, W09405. http://dx.doi.org/10.1029/2005WR004740

Deviney, F.A. (2009). Hierarchical continuous time Markov chain models for threshold exceedance at http://wwwlib.umi.com/dissertations/fullcit/3393526

DeWalle, D.R., Swistock, B.R., and Sharpe, W.E. (1995). Episodic flow-duration analysis: a method of assessing toxic exposure of brook trout (Salvelinus fontinalis) to episodic increases in aluminum. *Can. J. Fish. Aquat. Sci.,* 52(4), 816-827.

Gelman, A.B., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2004). *Bayesian Data Analysis*, CRC Press.

Gelman, A., and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*, Cambridge; New York, Cambridge University Press.

Hirsch, R.M., and Slack, J.R. (1984). Nonparametric Trend Test for Seasonal Data with Serial Dependence. *Water Resour. Res.,* 20(6), 727-732. http://dx.doi.org/10.1029/WR020i006p00727

Hirsch, R.M., Slack, J.R., and Smith, R.A. (1982). Techniques of Trend Analysis for Monthly Water Quality Data. *Water Resour. Res.,* 18(1), 107-121. http://dx.doi.org/10.1029/WR018i001p00107

Kocis, L., and Whiten, W.J. (1997). Computational investigations of low-discrepancy sequences. *ACM Trans. Math. Software,* 23(2), 266-294. http://dx.doi.org/10.1145/264029.264064

Laio, F., Porporato, A., Ridolfi, L., and Rodriguez-Iturbe, I. (2001). Plants in water-controlled ecosystems: active role in hydrologic processes and response to water stress II. Probabilistic soil moisture dynamics. *Adv. Water Resour.,* 24(7), 707-723. http://dx.doi.org/10.1016/S0309-1708(01)00005-7

Lu, Z., and Berliner, L.M. (1999). Markov switching time series models with application to a daily runoff series. *Water Resour. Res.,* 35(2), 523-534. http://dx.doi.org/10.1029/98WR02686

Morgan, M.M., and Grimshaw, A.S. (2007). Genesis II-Standards

Based Grid Computing, Seventh IEEE International Symposium on Cluster Computing and the Grid. http://dx.doi.org/10.1109/CCGRID.2007.53

Ross, S.M. (2006). *Introduction to Probability Models*, Academic Press.

Ryan, P.F., Hornberger, G.M., Cosby, B.J., Galloway, J.N., Webb, J.R., and Rastetter, E.B. (1989). Changes in the chemical composition of stream water in two catchments in the Shenandoah national park, Virginia, in response to atmospheric deposition of sulphur. *Water Resour. Res.,* 25(10), 2091-2099. http://dx.doi.org/10.1029/WR025i010p02091

Sickle, J.V., Baker, J., Simonin, H., Baldigo, B., Kretser, W., and Sharpe, W. (1996), Episodic Acidification of Small Streams in the Northeastern United States: Fish Mortality in Field Bioassays. *Ecol. Appl.,* 6(2), 408-421. http://dx.doi.org/10.2307/2269379

Skrondal, A., and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*, CRC Press. http://dx.doi.org/10.1201/9780203489437

Smith, R.L., Tawn, J.A., and Coles, S.G. (1997). Markov chain models for threshold exceedances. *Biometrika,* 84(2), 249-268. http://dx.doi.org/10.1093/biomet/84.2.249

Szilagyi, J., Balint, G., and Csik, A. (2006). Hybrid, Markov Chain-Based Model for Daily Streamflow Generation at Multiple Catchment Sites. *J. Hydrol. Eng.,* 11(3), 245. http://dx.doi.org/10.1061/(ASCE)1084-0699(2006)11:3(245)

Thomas, A., O'Hara, B., Ligges, U., and Sturtz, S. (2006). Making BUGS Open, *R News,* 6(1), 12-17.

Todorovic, P., and Zelenhasic, E. (1970). A Stochastic Model for Flood Analysis. *Water Resour. Res.,* 6(6), 1641-1648. http://dx.doi.org/10.1029/WR006i006p01641

Zhang, W., and Arhonditsis, G.B. (2008). Predicting the frequency of water quality standard violations using Bayesian calibration of eutrophication models. *J. Great Lakes Res.,* 34(4), 698-720. http://dx.doi.org/10.1016/S0380-1330(08)71612-5