Comparing corrosion control treatments for drinking water using a robust Bayesian generalized additive model

Benjamin F. Trueman§,*, Wendell James†, Trevor Shu†, Evelyne Doré§, and Graham A. Gagnon§

§Centre for Water Resources Studies, Department of Civil & Resource Engineering, Dalhousie University, 1360 Barrington St., Halifax, NS, CAN B3H 4R2
†EPCOR Water Services, Edmonton, AB, CAN

*Corresponding author: benjamin.trueman@dal.ca

Abstract

Pipe loop studies are used to evaluate corrosion control treatment, and updated regulatory guidance will ensure that they remain important for water quality management. But the data they generate are difficult to analyze: non-linear time-trends, non-detects, extreme values, and autocorrelation are common attributes that make popular methods, such as the t- or rank-sum tests, poor descriptive models. Here, we propose a model for pipe loop data that accommodates many of these difficult-to-model attributes: a robust Bayesian generalized additive model with continuous-time autoregressive errors. Our approach facilitates corrosion control treatment comparisons without the need for imputing non-detects or special handling of outliers. It is well-suited to describing nonlinear trends without overfitting, and it accounts for the reduced information content in autocorrelated time series. We demonstrate it using a four-year pipe loop study, with multiple pipe configurations and orthophosphate dosing protocols, finding that an initially high dose of orthophosphate (2 mg P L\(^{-1}\)) that is subsequently lowered (0.75 mg P L\(^{-1}\)) can yield lower lead release than an intermediate dose (1 mg P L\(^{-1}\)) in the long term. Water utilities face difficult tradeoffs in applying orthophosphate for corrosion control, and better models of pipe loop data can help inform the decision-making process.
Keywords

GAM; time series analysis; censored data; pipe loop; Lead and Copper Rule

Introduction

Pipe loop studies are used to model drinking water distribution systems, mainly for evaluating corrosion control treatment. The US EPA’s Revised Lead and Copper Rule mandates them under certain conditions, and the data they generate will need to be analyzed. But pipe loop study data are inherently complex and there is a real risk of misinterpreting them. In particular, (1) non-linearity, (2) censoring, (3) autocorrelation, (4) irregular sampling frequency, and (5) outliers are difficult to model using standard approaches (Figure 1a).

Non-linearity

Non-linear time trends and seasonal variation can be difficult to describe without over- or under-fitting model parameters. Nonlinearity can also make it difficult to identify a stable phase, as typical guidance for analyzing pipe loop study data suggests. The most common models for pipe loop data are usually inadequate to describe complex nonlinear variation.

Censoring

Censoring typically arises in pipe loop studies when analyte concentrations fall below the detection limit (i.e., left-censoring). Left-censoring is often addressed by substituting a constant for values below the detection limit and sometimes by removing the censored values. Neither approach is advisable as both introduce artificial patterns that can distort results, sometimes seriously. Popular statistical tests—for example, the paired t-test and its generalization, the one-way repeated measures ANOVA—cannot accommodate censoring. The non-parametric rank-sum test does allow for left-censoring at one detection limit, but it is not adequate for more complex censoring
schemes (e.g., multiple detection limits or interval censoring). Imputation methods such as regression on order statistics may be useful, but in a time series the censored data may be correlated, and imputation in the context of a full data model is usually preferable.

**Autocorrelation**

Autocorrelation means the current state of the system depends on its history. For example, the effect of a perturbation to corrosion scale is evident for some time after the fact. From a statistical perspective, autocorrelation makes parameter estimates more uncertain by reducing the effective sample size (i.e., the information content of the sample). As a consequence, it inflates the probability of falsely rejecting null hypotheses. Autocorrelation is very likely to be a feature of any pipe loop study dataset where effluent from pipes is sampled repeatedly over time, but it is not discussed in popular guides for pipe loop data analysis. Moreover, neither the t-test nor its non-parametric analogues account for autocorrelation explicitly.

The effect of autocorrelation can be surprisingly large: 100 observations from a time series may provide no greater precision than 10 independent observations if the 100 are strongly correlated in time. That is, $n$ autocorrelated observations provide less information than $n$ independent observations, and the difference can be large if autocorrelation is strong. Accounting for autocorrelation can improve predictive accuracy, but if it is ignored, analysts run the risk of making incorrect decisions based on the data.

**Irregular sampling frequency**

An irregular sampling frequency is common in pipe loop studies where various process interruptions preclude collecting a regular measurement series. Irregular sampling makes it more difficult to estimate the strength and structure of autocorrelation.
Outliers

Extreme lead or copper concentrations—typically due to particle release—are often inconsistent with the normality assumption common to many popular statistical tests. Outliers are sometimes omitted in practice, but unless they represent measurement error or a separate population (e.g., a process error), it is usually preferable to use a statistical method that is robust against them.

**Figure 1.** (a) An example time series from the study described in this paper; labels 1–5 denote attributes of the series presenting statistical modeling challenges. (b) Summary of the five components that comprise the proposed model.
A Bayesian approach

The primary goal of a pipe loop study is to compare two or more corrosion control treatments. But the data attributes outlined above can make it difficult to choose a statistical method from the suite that is commonly applied, given the necessary assumptions. Often, a more flexible approach is required, and Bayesian methods are inherently flexible since there is no need to derive the sampling distribution of an estimator.

Here, we propose a Bayesian generalized additive model for pipe loop study data that is sufficiently flexible for most pipe loop datasets. The collection of measurement series representing variation in a given parameter over time—often lead—is represented as the sum of five components (Figure 1b):

1. An average representing each pipe (all other components are centered on zero).
2. A smooth seasonal trend (i.e., a function of ordinal day), whose strength depends on the magnitude of the seasonal variation in water quality and temperature. This component may not be necessary if there is no distinct seasonality, or if the series is too short to identify it.
3. A smooth global trend describing multi-year variation in the parameter of interest across all pipes (i.e., a function of time).
4. A set of smooth pipe-specific trends that describe the deviations from the global multi-year trend particular to each pipe.
5. An autocorrelated error term.

Separation of the time-trend into multi-year and seasonal components allows for easier identification of a stable phase in the presence of significant seasonality, and separation of the multi-year trend into global and pipe-specific components facilitates trend estimation in the presence of censoring or where there are long data gaps in the series representing one or more pipes. In these instances, the model “borrows” information from the other time series. The autocorrelated error model improves the accuracy of parameter distributions and predictions while helping to prevent overfitting. And while time is the only covariate in our model, additional predictors can easily be added as smooth or linear functions.

We demonstrate the model using the data from a four-year pipe loop study that included multiple pipe configurations and orthophosphate dosing protocols. Our model
accommodates censored observations without imputation and is robust against extreme analyte concentrations due to particle release—meaning that no special treatment of outliers is necessary.

We use the model to (1) compare corrosion control treatments in terms of dissolved and particulate lead, (2) compare their relative contributions to total lead release, and (3) identify periods of stable lead release by considering the rate of change in the global multi-year term.

An advantage of our approach is the ability to move beyond point estimates and intervals: Bayesian models yield the full posterior distribution of all parameters and all quantities derived from those parameters, including the predicted values. This allows for propagation of uncertainty through all stages of post-processing and analysis.\textsuperscript{13} Moreover, the use of additive smoothing splines permits a description of time series using multiple nonlinear components, which can yield a more nuanced understanding of differences in analyte concentrations over time.

**Materials and methods**

**Model distribution system**

The model system comprised three pipe arrays (Figure S1), each with five pipes:

1. A 9.4 m, 1/2" ID lead pipe (Pb #1).
2. A duplicate of (1), Pb #2.
3. A 3 m, 1/2" ID lead pipe coupled galvanically to a 6.4 m, 1/2" ID copper pipe.
4. A 9.4 m, 1/2" ID copper pipe with twenty-two 50/50 lead-tin solder joints (“Cu/Pb-Sn”) (not discussed here).
5. A 9.4 m, 1/2" nominal internal diameter (ID) copper pipe (not discussed here).

All plumbing was new at the time of installation, and lead pipes were acquired from Canada Metals Eastern (ASTM B29, 99.94% pure lead). Each array was preceded in the flow path by a 3.6 m, 3" ID ductile iron pipe, and system pressure was approximately 40 psi.
The model system was conditioned from Sept. 2017 to Mar. 2018 using treated water from the E. L. Smith Water Treatment Plant in Edmonton, AB, CAN. Afterward, each array received a different dose of phosphoric acid (Table S1): 0 mg P L\(^{-1}\) (increasing to 0.5 mg L\(^{-1}\) in Jan. 2019), 1.0 mg L\(^{-1}\), and 2.0 mg L\(^{-1}\) (decreasing to 0.75 mg L\(^{-1}\) in Apr. 2019). The latter—a high “passivation” dose followed by a lower “maintenance” dose—is discussed elsewhere as a potential corrosion control strategy.\(^{14}\)

Beginning in Nov. 2020, influent to the model system was partially dechlorinated by blending granular activated carbon filtrate with chlorinated water upstream of the pipe sections. This was done to facilitate a biofilm study not discussed here, and the resulting total chlorine concentrations are summarized in Table S2. In Nov. 2020, 30 cm sections of lead pipe were cut from the galvanic lead-copper pipes for supplementary analysis. These sections were replaced with braided PVC hoses.

Flow through the pipe sections was programmed to approximate typical use, and details are available in Table S3. Samples were collected weekly until Mar. 2021, after which they were collected every two weeks. At each sampling event, 1 L was collected from each pipe after a 6-hour stagnation period. Samples were collected at a flow rate of approximately 3 L min\(^{-1}\).

**Water quality analysis**

Lead, iron, aluminum, and manganese were measured by ICP-MS\(^{15}\) with method detection limits estimated at 0.2, 5, 5, and 2 µg L\(^{-1}\), respectively. An aliquot of each sample was filtered at 0.45 µm using a polytetrafluoroethylene (PTFE) filter and a vacuum filter apparatus. Losses of dissolved lead by adsorption to PTFE filters have been documented previously,\(^{16}\) which here would tend to bias dissolved concentrations low and particulate concentrations high. In hard waters such as the feedwater for the model system, though, this bias is likely to be much lower than in soft waters.\(^{16}\)

Median lead in full and partial lead pipe sections (i.e., “Pb #1–2” and “Pb-Cu”) was 16 (5.3–120) and 1.7 (0.3–12) µg L\(^{-1}\), respectively (10\(^{th}\)–90\(^{th}\) percentiles in parentheses).
Time series representing effluent from individual pipes ranged in length from 181 to 187 date-value pairs spanning 1455 days.

Temperature was measured using an infrared thermometer, total chlorine\(^\text{17}\) and orthophosphate\(^\text{18}\) were measured using a spectrophotometer, and pH was measured using a benchtop meter. Water quality—including alkalinity,\(^\text{19}\) true colour,\(^\text{20}\) conductivity,\(^\text{21}\) fluoride,\(^\text{22}\) total hardness,\(^\text{23}\) and turbidity\(^\text{24}\)—is summarized in Table S4 (all Standard Methods). Colour was quantified after filtration at 0.45 µm, using mixed cellulose ester membrane filters.

![Figure 2](image-url)

**Figure 2.** (a) Constant, linear, and generalized additive models (GAMs) fitted to an example (log-transformed) time series from the dataset, along with the residuals from each model. (b)**
Pearson correlation ($r$) between the residuals $\epsilon$ at time $t$ and $t-1$. (c) Pearson correlation between the residuals $\epsilon$ at time $t$ and $t-k$, where $k$ is the lag and ranges from 0–21. The grey shaded regions represent $\pm 1.96/\sqrt{n}$, or an approximate 95% confidence bound on the autocorrelation function of an uncorrelated series.

Model selection

Comparisons of corrosion control treatments are often aided by a statistical model, and typical choices in pipe loop studies include constant (e.g., t-test) and linear models. But when applied to an example time series from our dataset (Figure 2a), both of these models violate a key assumption: the independence of model residuals.

Lack of independence is apparent in Figure 2b: the residual of either model at time $t$, $\epsilon_t$, is correlated with the residual at time $t-1$, $\epsilon_{t-1}$. This can be extended to include the residual at time $t-k$, $\epsilon_{t-k}$, where $k = 1...21$, yielding the autocorrelation function (Figure 2c). Independent residuals yield autocorrelation functions that remain close to zero, whereas correlated residuals yield autocorrelation functions that deviate from zero in a regular pattern: decay or oscillation, for instance. We assume here that the example time series is evenly-spaced, but in this case, that is only an approximation, and we relax it in the next section.
Figure 3. The unweighted (cubic) basis functions of \( x \) on which \( \log([\text{Pb}]) \) is regressed, and the same functions after weighting by the model coefficients. The (centered) model fit, or the sum of the weighted basis functions (equation 2), is superimposed in black.

The generalized additive model (GAM) does better than the constant or even the linear model in accounting for autocorrelation in the data (Figure 2). This is because it accounts for non-linear patterns that the simpler models ignore.\(^{25,26}\) GAMs can also be thought of as linear models, but instead of regressing \( y \) on \( x \), we regress \( y \) on \textit{functions} of \( x \) (cubic functions, for example, as in Figure 3).

\[
(1) \quad y_t = \beta_0 + f(t) + \epsilon_t
\]

where \( \beta_0 \) is the intercept, \( f(t) \) is a smooth function of \( t = \text{time} \),

\[
(2) \quad f(t) = \sum_{i=1}^{n} s_i(t) \beta_i
\]

\( s_i(t) \) are the basis functions of \( t \) that comprise \( f(t) \), \( \beta_i \) are the model coefficients (i.e., the basis function weights), and \( \epsilon_t \) is the error term. We fitted the example GAM in

\( \sum_{i=1}^{n} s_i(t) \beta_i \)
Figures 2–3 using the R package mgcv \(^27\) which uses penalized likelihood maximization instead of ordinary least squares to choose coefficients while protecting against overfitting.

**Residual autocorrelation**

Often, even a flexible model—such as a GAM—fails to account for residual autocorrelation in a time series, in which case it is possible to relax the independence assumption and model the residuals as an *autoregressive* process. A very common model is the first-order autoregressive (AR(1)) process:

\[
\epsilon_t = \phi \epsilon_{t-1} + \omega_t
\]

where \(\epsilon_t\) is the residual at time \(t\), \(\phi\) is the autoregressive parameter, and \(\omega_t\) is an independent draw from a continuous—typically Gaussian—probability distribution.

First-order autocorrelation can be approximated as the correlation between the lagged and unlagged residuals, \(\epsilon_t\) and \(\epsilon_{t-1}\). A reasonable estimate of \(\phi\) for the linear model residuals in Figure 2b, then, would be \(r\). Unmodeled positive autocorrelation, as in Figure 2b-c, yields parameter estimates that are overly precise, and so it should be either eliminated via model selection or taken into account explicitly. Here, we chose to analyze the full length of each time series—including the conditioning period—to best estimate the autocorrelation structure.

And while equation (3) is limited to time series that are regularly observed, it can be generalized to an irregular sequence of observations as a continuous-time first-order autoregression.\(^4\) This more general model has the following autocorrelation function:

\[
h(s, \phi) = \phi^s, \quad s \geq 0, \quad 0 \leq \phi \leq 1
\]

In equation (4), \(s\) represents the separation in time between successive observations. Autocorrelation, then, decays exponentially in time; if two observations are spaced far enough apart, they are approximately independent. Given the irregular spacing of
observations in our data, we modeled autocorrelation using the function defined in equation (4).

**Model description**

To compare corrosion control treatments using data from the model distribution system, we fitted separate GAMs with continuous-time first-order autocorrelated errors to the particulate (>0.45 µm) and dissolved (<0.45 µm) lead time series. Particulate lead was determined as the difference between total and dissolved lead. Particulate lead concentrations less than the detection limit of 0.2 µg L\(^{-1}\) were left-censored, and particulate lead concentrations representing the difference between an observed (total) and left-censored (dissolved) concentration were interval-censored. The upper bound of the interval was set as the total lead concentration, implying zero dissolved lead. The lower bound was set as the total lead concentration less the detection limit, implying a dissolved lead concentration equal to the detection limit. Given their impractically high censoring rates (up to 96%), time series representing effluent from soldered copper pipe sections were not modeled.

Model code was written in Stan\(^{28,29}\), a C++ package for Bayesian inference using Markov chain Monte Carlo (MCMC) methods. We generated a code template using the R package \textit{brms}\(^{30,31}\) and modified it to specify the continuous-time autocorrelation structure defined in equation (4). We used a suite of R packages,\(^{32-43}\) and bespoke functions used to fit the model and process the results are documented on GitHub.\(^{44}\) The data and R code necessary to reproduce the analysis are available in a separate repository.\(^{45}\)
Figure 4. (a) Seasonal component of the trend in dissolved lead. (b) Global multi-year component of the trend in dissolved and particulate lead. (c) Group-level (i.e., pipe-specific) multi-year trends in dissolved and particulate lead, representing the deviations of each time series from the global multi-year trend. In panels (a–b), shaded regions span the middle 95% of the posterior distribution, and 200 draws from that posterior are superimposed in blue. All plots are shown on the transformed scale and centered.

The models included one-dimensional global smoothers describing multi-year variation in lead across all time series (Figure 4b). The model of dissolved lead also included a global smoother describing seasonal variation (Figure 4a). And while the model assumes that seasonality strength does not vary over time, some previous work has suggested that orthophosphate may damp seasonal variation in lead. This effect could be incorporated by allowing the seasonal smoother to vary with the orthophosphate dose.

Both models included group-level smoothers describing multi-year variation in lead release from each pipe (Figure 4c), a hierarchical structure that is discussed elsewhere. The multi-year smoothers were thin-plate regression splines and the
seasonal smoother was a cyclic cubic regression spline.\textsuperscript{27} To speed up computation, the response variable was log-transformed and then scaled to have zero mean and unit variance. And while it is omitted here, total lead could be predicted using the same model structure as dissolved lead.

Time series representing lead release to drinking water tend to feature extreme values caused by particulate or colloidal lead. Even in 0.45 $\mu$m filtrate (i.e., dissolved lead), small colloids can sometimes elevate lead concentrations well above solubility limits.\textsuperscript{48} These “outliers” can severely bias parameter estimates unless they are accounted for. And since our dataset was impacted by extreme lead release events, we used a Student $t$ likelihood in place of the usual Gaussian likelihood to limit the influence of extreme values.\textsuperscript{30,31}

Applying Bayes’ theorem,\textsuperscript{49} the joint probability distribution of the model parameters, conditional on the data (i.e., the posterior), is proportional to product of the likelihood—the probability of the data, conditional on the model parameters—and the prior (i.e., the probability distribution of the model parameters uninformed by the data).

The transformed dissolved or particulate lead time series, $y_t$, were modeled as follows:

$$
\begin{align*}
    &\text{likelihood:} \\
    &y_t|\text{censored}_t = 0 \sim T(\mu_t, \sigma, \nu) \quad \text{[observed]} \\
    &y_t|\text{censored}_t = 1 \sim F(\mu_t, \sigma, \nu) \quad \text{[left-censored]} \\
    &y_t|\text{censored}_t = 2 \sim g(\mu_t, \sigma, \nu) \quad \text{[interval-censored]}
\end{align*}
$$

model for $\mu_t$:

$$
\begin{align*}
    \mu_t &= \alpha_{\text{pipe}t} + \sum_{j=1}^{n} f_j(t) + \phi^s r_{t-s} \\
    r_{t-s} &= y_{t-s} - \alpha_{\text{pipe}t} - \sum_{j=1}^{n} f_j(t - s)
\end{align*}
$$

priors:

$\sigma \sim \text{Half-T}(0,2.5,3)$
$\nu \sim \text{Gamma}(2,0.5)$
$\phi \sim N(0.5,0.25)$
where $T$ represents the Student $t$-distribution with mean $\mu_t$, standard deviation $\sigma$, and degrees-of-freedom parameter $v$, determining the difference between the $t$-distribution and a Gaussian with the same $\mu$ and $\sigma$. ($\text{Gamma}$ and $N$ represent the gamma and Gaussian distributions, parameterized by the mean and shape parameter/standard deviation, respectively.)

Left-censored observations are indicated by the variable $\text{censored} = 1$ and were modeled using $F_X(l)$, the cumulative Student $t$-distribution ($i.e.$, $P(X \leq l)$, the probability that, given a set of distributional parameters, a random variable $X$ is equal to the censoring limit, $l$, at most). Interval-censored observations are indicated by $\text{censored} = 2$ and were modeled using $g(l, u)$, or $P(l < X \leq u) = F_X(u) - F_X(l)$, where $l$ and $u$ are the lower and upper limits, respectively. Parameters, then, were estimated without imputing nondetects: at each iteration of the MCMC algorithm, censored observations factor into the computed posterior via the cumulative probability distribution, $F_X(x)$, only.

The group-level intercepts $\alpha_{\text{pipe}_i}$ (i.e., the pipe-specific average responses) are defined as follows:

\[
\alpha_{\text{pipe}_i} \sim N(\alpha, \sigma_\alpha) \quad \text{[adaptive prior]} \\
\alpha \sim T(0,1.2,5,3) \quad \text{[prior on global intercept]} \\
\sigma_\alpha \sim \text{Half-T}(0,2.5,3) \quad \text{[prior on } \sigma \text{ of pipe averages]}
\]

The autocorrelation coefficient $\phi^5$ is defined in equation (4), and $\tau_{t-s}$ is the GAM residual at the previous time step, $t - s$. The $f_j(t)$ are smooth spline functions of time that comprise the GAM, and each of the $n$ smoothers takes the following form:

\[
\text{generalized additive model (GAM):} \\
f_j(t) = X_j \beta_j + Z_j b_j
\]

\[
\text{priors on GAM parameters:} \\
\beta_j \sim T(0,2.5,3) \quad \text{[unpenalized]} \\
b_j \sim N(0, \sigma_b) \quad \text{[penalized]} \\
\sigma_b \sim \text{Half-N}(0,0.5) \quad \text{[seasonal smoother]} \\
\sigma_b \sim \text{Half-T}(0,2.5,3) \quad \text{[global/group-level smoothers]}
\]
where $Z_j$ and $X_j$ are matrices representing the penalized and unpenalized basis functions evaluated at time $t$; and $\beta$ and $b_j$ represent the corresponding spline coefficient vectors ($\beta$ is scalar, corresponding to a single unpenalized basis function included in the global multi-year trend). The priors on $\sigma_b$ help to prevent overfitting by penalizing complexity in the fitted smooth terms.\(^{27}\) And while the chosen priors did help with model convergence, they did not exert a strong influence on the posterior (Figure S2); the likelihood—representing 1663 observations in each model—was overwhelmingly influential. A summary of the fitted models, including convergence statistics and effective sample sizes, is provided in Figures S3 and S4.

**Estimating rates of change**

We estimated the first derivative of the global multi-year smoothers using a finite difference method. Specifically, we generated 4000 posterior predictions of the smooth trend along a regular sequence spanning the date range of the data. This was repeated after adding $\delta = 0.001$ (~1/3 of a day) to each value in the sequence. The difference between the grid of posterior predictions evaluated at $t$ and $t + \delta$, divided by $\delta$, approximates the instantaneous rate of change after accounting for seasonal variation, inter-pipe variation, and autocorrelation.

**Model parameters and diagnostics**

Residuals from the two models were approximated well by the estimated residual error distribution (Figure 5a). Furthermore, models fitted to a draw from the posterior predictive distribution—that is, to simulated observations—yielded similar parameters to those of the original models (Figures S5–7).
Figure 5. (a) Histograms representing the median of 4000 residual draws from the models; the theoretical densities (red curves) are calculated from the posterior medians of the distributional parameters (i.e., \( \sigma, \nu \)) from each model. (b) Density of first-order autocorrelation, estimated using the Spearman rank correlation coefficient and 4000 residual draws from the model including and excluding the continuous-time first-order autocorrelation term (CAR(1) and GAM, respectively). Estimates were generated using an augmented dataset where censored observations were imputed via posterior prediction. (c) Posterior distributions of \( \phi \), the autocorrelation coefficient. (d) Pareto \( k \) shape parameters, which estimate the influence of each observation on the posterior distribution.

The continuous-time autocorrelated error term was successful in capturing residual first-order autocorrelation, estimated in Figure 5b using the Spearman rank correlation coefficient as a measure that is robust against outliers. Since the difference between
model predictions and the censoring limit is not an accurate estimate of residual model error, we simulated residuals for model assessment by imputing censored values with posterior predictions and computing the residuals using the augmented dataset. Simulated residuals were minimally autocorrelated, while predictions that excluded the autocorrelation term yielded much stronger residual autocorrelation. The distribution of the autocorrelation coefficient for each model, \( \phi \), is shown in Figure 5c, and simulated residuals as a function of time are shown in Figure S8.

The posterior predictive distribution was also robust against extreme values (Figure 5d): the Pareto-\( k \) values shown here represent the influence of individual observations on the posterior distribution, and values below 0.7 are considered moderately influential (at most), yielding reliable estimates of out-of-sample predictive accuracy (see Comparing corrosion control treatments below).\(^{50,51}\) Here, all Pareto-\( k \) values were below 0.5.

## Results and discussion

### Comparing corrosion control treatments

Time series of particulate and dissolved lead featured prominent negative trends after the introduction of orthophosphate, followed by an extended period of relative stability (Figure 6). The dissolved lead time series were seasonal, with maximum levels in late summer and minimum levels in early spring (Figure 4a).
Figure 6. Time series of dissolved and particulate lead representing three orthophosphate dosing protocols (columns) and two pipe configurations (rows; the full lead pipe configuration—Pb #1–2—is duplicated). The expected value of the posterior predictive distribution (i.e., $\mu_t$ in Equation 5) is summarized by the median (blue and red lines) and the 2.5–97.5th percentiles (red/blue shaded regions). The yellow vertical bands represent brief periods when orthophosphate dosing was interrupted, and ticks at the top of each panel represent values outside the plot limits. Orthophosphate doses are annotated.

Posterior predictions from the models—$\mu_t$ in Equation (5)—can be compared to evaluate the orthophosphate dosing protocols. Here we compared the 0–0.5 mg P L$^{-1}$ with the 1 mg P L$^{-1}$ and the 1 mg P L$^{-1}$ with the 2–0.75 mg P L$^{-1}$ dosing protocols over time by considering their ratios. This highlights an advantage of the Bayesian approach—straightforward error propagation for quantities derived from the posterior distribution of the model parameters.52

The three dosing protocols yielded distinct lead release distributions over most of the study, and the difference between the lowest orthophosphate dose and the other two is particularly apparent (Figure 7). The ratio of dissolved Pb concentrations at a P ratio of 0:1—comparing the 0 mg P L$^{-1}$ control with the 1 mg P L$^{-1}$ dose—peaked at 31.3 with a
95% credible interval of 27.3–35.5 (Figure 7a). That is, lead release at 0 mg P L\(^{-1}\) was as much as 31 times higher than lead release at 1 mg P L\(^{-1}\).

The dissolved Pb ratio declined following the introduction of orthophosphate at 0.5 mg P L\(^{-1}\) (i.e., at a 0.5:1 P ratio), but its 2.5\(^{th}\) percentile—the lower bound of the 95% credible interval—remained above 1 with a single brief exception. This highlights the superior corrosion control performance of 1 mg P L\(^{-1}\) over both 0 and 0.5 mg P L\(^{-1}\) and is consistent with expectations based on geochemical solubility models.\(^{53}\)

Pb ratios comparing the 1 and 2–0.75 mg P L\(^{-1}\) protocols exhibited a more nuanced pattern (Figure 7b). As a comparison of either particulate or dissolved lead concentrations, the Pb ratio at a 1:2 P ratio was less than one, meaning that 2 mg P L\(^{-1}\) caused excess lead release relative to the 1 mg P L\(^{-1}\) dose. This likely occurred in the dissolved fraction as small colloids.\(^{54–56}\)

The transition from 2–0.75 mg P L\(^{-1}\) reversed this pattern: the dissolved and particulate Pb ratios at the 1:0.75 P ratio were largely greater than one, meaning greater lead release at the 1 mg P L\(^{-1}\) dose. A high dose of 2 mg P L\(^{-1}\), then, presumably caused a scale layer to form that ultimately released less lead—in both fractions—than its counterpart formed at 1 mg P L\(^{-1}\). This is despite being exposed to a lower orthophosphate concentration in the latter half of the study (0.75 mg P L\(^{-1}\)), and it highlights a potential advantage and drawback of the passivation-maintenance dosing strategy. That is, a high passivation dose may cause excess lead release in the short term—likely due to particle-generating mechanisms\(^{56,57}\)—while ensuring that a lower maintenance dose is either more effective or effective sooner than it otherwise might be.
Figure 7. Ratios of posterior predictions comparing lead concentrations at (a) 0–0.5 with 1 mg P L\(^{-1}\) (numerator and denominator, respectively) and (b) 1 with 2–0.75 mg P L\(^{-1}\). The bold black/red line represents the median of the posterior distribution and the shaded grey region spans the 95% credible interval; black and red indicate portions of the series where the interval included and excluded, respectively, a ratio of 1 (i.e., equality). A random sample of 200 posterior draws is superimposed in blue.

**Identifying stability**

In a typical pipe loop study, effluent lead concentrations decrease sharply after commissioning. Assuming inputs to the system remain constant, a period of relative stability usually follows.\(^3\) This pattern is represented in our model by the global multi-year trend (Figure 4b), which excludes pipe-specific and seasonal water quality variation. The stability of a pipe array—accounting for effluent from all pipes in the study—can be evaluated by considering the rate of change in this global term. Here,
dissolved and particulate lead concentrations in pipe loop effluent declined significantly in the first year of operation as the system stabilized (Figure 8a). The subtler subsequent positive and negative trends in both fractions are difficult to explain, but they may be a response to changes in water quality not captured directly by the model (see Trends in particulate lead release).

**Figure 8.** (a) The global multi-year smoother ($f_{\text{global}}(t)$) and its estimated first derivative ($f'_{\text{global}}(t)$). A random sample of 200 curves/slope estimates is superimposed, and both sets of lines are underlain by shaded regions covering the middle 95% of the respective posterior distributions. The median of the global trend is shaded red when the 95% credible interval around $f'_{\text{global}}(t)$ does not include zero. (b–c) Iron and water temperature, both measured in pipe rack influent,
exhibited comparable seasonal patterns to lead, the latter described by the seasonal term in the 
GAM. Ticks at the top and bottom edges of the plot in (b) represent anomalously high and left-
censored iron concentrations, respectively.

The variation in dissolved lead evident in Figure 8a neglects seasonality by design. And 
while the seasonal component of the model (Figure 8b) is described as a function of 
time (ordinal day), this is a proxy for the physical and chemical inputs that are expected 
to have caused it. The true seasonal relationship is probably complex and difficult to 
determine from the study data, but temperature and influent iron concentration are 
plausible drivers: both varied seasonally and have been shown to impact lead release 
(Figure 8b–c).58–61 And although dissolved organic carbon has also been shown to 
increase lead solubility, 48,62,63 true colour—a proxy for dissolved organic carbon64—was 
not strongly seasonal (Figure S9).

**Trends in particulate lead release**

Over much of the study, the dissolved fraction (<0.45 µm) dominated total lead release 
(Figure 6). During the conditioning phase in particular—before orthophosphate was 
introduced—particles represented a minor contribution. Orthophosphate treatment 
caus...
Modeling autocorrelation

Autocorrelation models the dependence of observations on previous observations, or the response of a measurement series at time \( t + s \) to a shock at time \( t \). The effect of the autocorrelated error term here is most evident in the response of the model to high particulate lead release after short sections of lead pipe were removed from the lead-copper pipe couples in Nov. 2020 (“Pb-Cu”, Figure 6). While the time-varying mean \( \mu_t \) exhibits pronounced spikes corresponding to particle release events, the underlying GAM remains smooth (Figure 4). That is, extreme and sustained deviations from typical behavior are generally captured by the error term (but still reflected in model predictions: see Figure 6 and equation 5).

Accounting for autocorrelation improved the predictive accuracy of the models, assessed here using approximate leave-one-out cross-validation on the full dataset. Approximate leave-one-out cross-validation estimates the expected log pointwise predictive density—a metric of model accuracy—for new observations not used to fit the model.\(^{49–51}\) We applied it to the particulate and dissolved lead models with continuous-time autoregressive errors and also to identically-specified models with independent errors. The autoregressive error models yielded higher expected log pointwise predictive densities (i.e., greater accuracy), exceeding the independent error models in this metric by 206 and 47 (dissolved and particulate lead models, respectively).

Conclusion

Pipe loop study data can be difficult to analyze, owing to non-linear time trends, seasonality, autocorrelation due to repeated measurements of pipe effluent, irregular sampling frequency, extreme particulate metals concentrations, and left-censoring, or values below the detection limit. To address these difficulties, we propose a robust Bayesian generalized additive model with continuous-time first-order autoregressive errors. With this approach, there is no need to impute non-detects, assume linearity, or
treat outliers specially. Moreover, the model accounts for the reduced information content in autocorrelated observations.

Our approach can be modified to suit the needs of diverse analyses. While time is the sole independent variable here, water quality parameters can easily be added as predictors if inferences concerning their effects are of interest. Moreover, any parameter in the likelihood can be modeled—including the standard deviation if model residuals are heteroskedastic. The model is equally useful for other problems featuring the partially-censored, irregularly-sampled, autocorrelated time series common in water quality analysis. But like many Bayesian models, this one is computationally-intensive to fit and would be difficult to implement in real time.

We demonstrate it here using data from a four-year pipe loop study and we use the model to compare orthophosphate dosing protocols and analyze time trends. The additive smooth terms comprising the model yield an informative decomposition of pipe loop time series and a more nuanced comparison of corrosion control treatments.

The pipe loop system responded strongly to orthophosphate treatment, as evident in the decreasing global trend during the first year of operation and the large difference in lead release between the 0 and 1 mg P L\(^{-1}\) dose. But the effect of orthophosphate dose was complex: lead release at 1 mg P L\(^{-1}\) was lower than at 0.5 mg P L\(^{-1}\) but higher than at 0.75 mg P L\(^{-1}\). We attribute the superior performance of the 0.75 mg P L\(^{-1}\) dose to an initial high dose of 2 mg P L\(^{-1}\). On the other hand, the initial 2 mg P L\(^{-1}\) dose caused excess lead release, highlighting an important trade-off that should be considered in applying a passivation-maintenance (high-low) dosing strategy. And since high orthophosphate doses may generate small lead-orthophosphate colloids/nanoparticles that are difficult to remove from drinking water,\(^{56}\) the 1 mg P L\(^{-1}\) dose may be the better choice. The additional cost and downstream effects of elevated orthophosphate doses should also be considered: even in light of a thorough pipe loop study, water utilities face difficult decisions in applying orthophosphate.
Given that pipe loop studies are expected to remain important for drinking water quality management, time series methods will be necessary for reliable statistical analysis of the data they generate. Methods that do not account for the effect of repeated measurements—and the reduced information content that results from autocorrelation—are likely to be less useful, and may even be misleading. Methods should be flexible enough to account for nonlinear time trends, and they should be robust against extreme values. And since the goal of pipe loop studies is often to minimize the release of lead and other metals into drinking water, concentrations may fall below the analytical range for a given analyte. The most useful statistical methods, then, will be well-suited to modeling censored data.

**Acknowledgements**

This work was supported by Mitacs, through the Mitacs Accelerate Program (Reference # IT23352), and NSERC, through an Industrial Research Chair program (Grant # IRCPJ: 349838-16) and a Postdoctoral Fellowship (E. Doré). We thank David Redden and Paul Bjorndahl for providing comments on a draft of the manuscript. We also acknowledge EPCOR’s analytical operations team for water quality analysis and their process development team for operating the pipe loops.

**Associated content**

**Supporting information**

The Supporting Information is available free of charge via the internet at [http://pubs.acs.org](http://pubs.acs.org).

Additional description of the pipe loop study and the statistical model, including model summaries, convergence statistics, and a comparison of different prior choices.
References


(33) Lee, L. *NADA: Nondetects and data analysis for environmental data*. https://CRAN.R-project.org/package=NADA.


(41) van den Brand, T. *gg4x: Hacks for ggplot2*. https://CRAN.R-project.org/package=gg4x.


**TOC graphic**
Supplementary information for Comparing corrosion control treatments for drinking water using a robust Bayesian generalized additive model

Benjamin F. Trueman§,*, Wendell James†, Trevor Shu†, Evelyne Doré§, and Graham A. Gagnon§

§Centre for Water Resources Studies, Department of Civil & Resource Engineering, Dalhousie University, 1360 Barrington St., Halifax, NS, CAN B3H 4R2

†EPCOR Water Services, Edmonton, AB, CAN

*Corresponding author: benjamin.trueman@dal.ca

This document contains 9 figures, 4 tables, and 13 pages.
**Table S1.** Orthophosphate dosing schedule.

<table>
<thead>
<tr>
<th>Array</th>
<th>Ortho-P Dosing (mg/L, as P)</th>
<th>Timeline</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Conditioning, 0.0</td>
<td>Sep. 27, 2017 - Jan. 29, 2019</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>Jan. 29, 2019 onward, to date</td>
</tr>
<tr>
<td>2</td>
<td>Conditioning, 0.0</td>
<td>Sep. 27, 2017 - Mar. 13, 2018</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Mar. 13, 2018 onward, to date</td>
</tr>
<tr>
<td>3</td>
<td>Conditioning, 0.0</td>
<td>Sep. 27, 2017 - Mar. 13, 2018</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Mar. 13, 2018 - Apr. 16, 2019</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>Apr. 16, 2019 onward, to date</td>
</tr>
</tbody>
</table>

**Table S2.** Total chlorine in pipe section influent (mg L⁻¹).

<table>
<thead>
<tr>
<th>Period</th>
<th>Mean</th>
<th>St. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before Dec. 14, 2020</td>
<td>2.120</td>
<td>0.210</td>
</tr>
<tr>
<td>Dec. 14, 2020 – Mar. 30, 2021</td>
<td>0.648</td>
<td>0.121</td>
</tr>
<tr>
<td>Mar. 31, 2021 - June 4, 2021</td>
<td>0.093</td>
<td>0.056</td>
</tr>
<tr>
<td>June 8 – Nov. 29, 2021</td>
<td>0.282</td>
<td>0.049</td>
</tr>
</tbody>
</table>
Table S3. Pipe array flow schedule; sampling occurred at T4, and FY-001, -002, and -003 are defined in Figure S1. A Modicon M221 logic controller (Schneider Electric, Canada) was used for process control. Three 2-inch ASCO 8210G100 solenoid valves (i.e., FY-001, -002 and -003) were wired to the logic controller, opening and closing according to the schedule.

<table>
<thead>
<tr>
<th>Time point</th>
<th>Actual time (daylight time)</th>
<th>Actual time (summer time)</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>T0</td>
<td>0200 hrs</td>
<td>0300 hrs</td>
<td>FY-001, 002, 003 open</td>
</tr>
<tr>
<td>T1</td>
<td>0229 hrs 50 secs</td>
<td>0329 hrs 50 secs</td>
<td>FY-001 close</td>
</tr>
<tr>
<td>T2</td>
<td>0230 hrs</td>
<td>0330 hrs</td>
<td>FY-002 close</td>
</tr>
<tr>
<td>T3</td>
<td>0230 hrs 10 secs</td>
<td>0330 hrs 10 secs</td>
<td>FY-003 close</td>
</tr>
<tr>
<td>T4</td>
<td>0830 hrs</td>
<td>0930 hrs</td>
<td>FY-001, 002, 003 open</td>
</tr>
<tr>
<td>T5</td>
<td>0834 hrs 50 secs</td>
<td>0934 hrs 50 secs</td>
<td>FY-001 close</td>
</tr>
<tr>
<td>T6</td>
<td>0835 hrs</td>
<td>0935 hrs</td>
<td>FY-002 close</td>
</tr>
<tr>
<td>T7</td>
<td>0835 hrs 10 secs</td>
<td>0935 hrs 10 secs</td>
<td>FY-003 close</td>
</tr>
<tr>
<td>T8</td>
<td>0915 hrs</td>
<td>1015 hrs</td>
<td>FY-001, 002, 003 open</td>
</tr>
<tr>
<td>T9</td>
<td>1014 hrs 50 secs</td>
<td>1114 hrs 50 secs</td>
<td>FY-001 close</td>
</tr>
<tr>
<td>T10</td>
<td>1015 hrs</td>
<td>115 hrs</td>
<td>FY-002 close</td>
</tr>
<tr>
<td>T11</td>
<td>1015 hrs 10 secs</td>
<td>1115 hrs 10 secs</td>
<td>FY-003 close</td>
</tr>
<tr>
<td>T12</td>
<td>1230 hrs</td>
<td>1330 hrs</td>
<td>FY-001, 002, 003 open</td>
</tr>
<tr>
<td>T13</td>
<td>1329 hrs 50 secs</td>
<td>1429 hrs 50 secs</td>
<td>FY-001 close</td>
</tr>
<tr>
<td>T14</td>
<td>1330 hrs</td>
<td>1430 hrs</td>
<td>FY-002 close</td>
</tr>
<tr>
<td>T15</td>
<td>1330 hrs 10 secs</td>
<td>1430 hrs 10 secs</td>
<td>FY-003 close</td>
</tr>
<tr>
<td>T16</td>
<td>1730 hrs</td>
<td>1830 hrs</td>
<td>FY-001, 002, 003 open</td>
</tr>
<tr>
<td>T17</td>
<td>1929 hrs 50 secs</td>
<td>2029 hrs 50 secs</td>
<td>FY-001 close</td>
</tr>
<tr>
<td>T18</td>
<td>1930 hrs</td>
<td>2030 hrs</td>
<td>FY-002 close</td>
</tr>
<tr>
<td>T19</td>
<td>1930 hrs 10 sec</td>
<td>2030 hrs 10 sec</td>
<td>FY-003 close</td>
</tr>
<tr>
<td>T20</td>
<td>2330 hrs</td>
<td>0030 hrs</td>
<td>FY-001, 002, 003 open</td>
</tr>
<tr>
<td>T21</td>
<td>2359 hrs 50 secs</td>
<td>0059 hrs 50 secs</td>
<td>FY-001 close</td>
</tr>
<tr>
<td>T22</td>
<td>0000 hrs</td>
<td>0100 hrs</td>
<td>FY-002 close</td>
</tr>
<tr>
<td>T23</td>
<td>0000 hrs 10 secs</td>
<td>0100 hrs 10 secs</td>
<td>FY-003 close</td>
</tr>
</tbody>
</table>
Table S4. Water quality summary.

<table>
<thead>
<tr>
<th>Analyte</th>
<th>Units</th>
<th>Location</th>
<th>Period</th>
<th>Median</th>
<th>1st quartile</th>
<th>3rd quartile</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alkalinity</td>
<td>mg/L</td>
<td>Treated</td>
<td>Comprehensive</td>
<td>123.0000</td>
<td>114.0000</td>
<td>131.0000</td>
<td>1520</td>
</tr>
<tr>
<td>Aluminum</td>
<td>mg/L</td>
<td>LP[1-3]A</td>
<td>Comprehensive</td>
<td>0.0490</td>
<td>0.0270</td>
<td>0.081</td>
<td>549</td>
</tr>
<tr>
<td>Colour</td>
<td>TCU</td>
<td>Treated</td>
<td>Comprehensive</td>
<td>0.8930</td>
<td>0.6710</td>
<td>1.160</td>
<td>1520</td>
</tr>
<tr>
<td>Conductivity</td>
<td>µmho</td>
<td>Treated</td>
<td>Comprehensive</td>
<td>387.0000</td>
<td>367.0000</td>
<td>408.0000</td>
<td>1520</td>
</tr>
<tr>
<td>Fluoride</td>
<td>mg/L</td>
<td>Treated</td>
<td>Comprehensive</td>
<td>0.6610</td>
<td>0.1200</td>
<td>0.690</td>
<td>1520</td>
</tr>
<tr>
<td>Manganese</td>
<td>mg/L</td>
<td>LP[1-3]A</td>
<td>Comprehensive</td>
<td>0.0050</td>
<td>0.0030</td>
<td>0.007</td>
<td>551</td>
</tr>
<tr>
<td>Ortho-P Residual</td>
<td>mg/L as P</td>
<td>LP1A</td>
<td>Comprehensive</td>
<td>0.4900</td>
<td>0.4600</td>
<td>0.520</td>
<td>142</td>
</tr>
<tr>
<td>Ortho-P Residual</td>
<td>mg/L as P</td>
<td>LP2A</td>
<td>Comprehensive</td>
<td>0.9400</td>
<td>0.8700</td>
<td>1.000</td>
<td>185</td>
</tr>
<tr>
<td>Ortho-P Residual</td>
<td>mg/L as P</td>
<td>LP3A</td>
<td>Dose 1</td>
<td>1.8900</td>
<td>1.7200</td>
<td>1.960</td>
<td>54</td>
</tr>
<tr>
<td>Ortho-P Residual</td>
<td>mg/L as P</td>
<td>LP3A</td>
<td>Dose 2</td>
<td>0.7300</td>
<td>0.6800</td>
<td>0.760</td>
<td>130</td>
</tr>
<tr>
<td>Total Hardness</td>
<td>mg/L</td>
<td>Treated</td>
<td>Comprehensive</td>
<td>169.0000</td>
<td>161.0000</td>
<td>171.0000</td>
<td>1230</td>
</tr>
<tr>
<td>Turbidity</td>
<td>NTU</td>
<td>Treated</td>
<td>Comprehensive</td>
<td>0.0529</td>
<td>0.0456</td>
<td>0.060</td>
<td>1520</td>
</tr>
</tbody>
</table>
Figure S1. Schematic of the pipe loop system.
Figure S2. Choice of prior did not appear to have a strong impact on the posterior distribution. Figures S3–4 and S6–7 summarize the parameters of models fitted to observed and simulated data, respectively, including all of the spline coefficients (the weights applied to the basis functions). The statistic $\hat{R}$ is a convergence metric; at convergence $\hat{R}$ is equal to 1, and an $\hat{R}$ less than 1.1 is usually taken to represent approximate convergence\(^1\). Here, all $\hat{R}$ values were less than 1.05. The effective sample size $N_{eff}$ measures the amount by which autocorrelation in the MCMC chains increases uncertainty in the parameter estimates. That is, $N_{eff}$ has the same estimation power as the $N$ autocorrelated samples. An effective sample size greater than 100 per chain is considered adequate. Here, effective sample sizes for all parameters exceeded this threshold.
Figure S3. Summary of model parameters and convergence statistics (dissolved lead model). Medians calculated from the posterior distributions of the (a) spline coefficients, (b) other parameters (as labelled), (c) random intercepts, and (d) standard deviations of the group-level multiyear splines. Error bars span the middle 95% of each distribution. $\hat{R}$, a convergence statistic, is mapped to point colour and size.
Figure S4. Summary of model parameters and convergence statistics (particulate lead model). Medians calculated from the posterior distributions of the (a) spline coefficients, (b) other parameters (as labelled), (c) random intercepts, and (d) standard deviations of the group-level multiyear splines. Error bars span the middle 95% of each distribution. $\hat{R}$, a convergence statistic, is mapped to point colour and size.
Figure S5. In the model fitted to simulated data, (a) seasonal component of the trend in dissolved lead. (b) Global multiyear component of the trend in dissolved and particulate lead. (c) Group-level (i.e., pipe-specific) multiyear trends in dissolved and particulate lead, representing the deviations of each time series from the global multiyear trend. In panels (a–b), shaded regions span the middle 95% of the posterior distribution, and 200 draws from that posterior are superimposed in blue. All plots are shown on the transformed scale and centered.
Figure S6. Summary of model parameters and convergence statistics (dissolved lead model fitted to simulated data). Medians calculated from the posterior distributions of the (a) spline coefficients, (b) other parameters (as labelled), (c) random intercepts, and (d) standard deviations of the group-level multiyear splines. Error bars span the middle 95% of each distribution. $\hat{R}$, a convergence statistic, is mapped to point colour and size.
Figure S7. Summary of model parameters and convergence statistics (particulate lead model fitted to simulated data). Medians calculated from the posterior distributions of the (a) spline coefficients, (b) other parameters (as labelled), (c) random intercepts, and (d) standard deviations of the group-level multiyear splines. Error bars span the middle 95% of each distribution. $\hat{R}$, a convergence statistic, is mapped to point colour and size.
Figure S8. Simulated residuals, grouped by orthophosphate dosing protocol and pipe material. Points represent medians summarizing 4000 draws from the posterior, and error bars span the 95% credible intervals.
Figure S9. (a) True colour in pipe loop feedwater, by month and year. (b) Aluminum in feedwater, by orthophosphate dose.

References