River Water-Quality Concentration and Flux Estimation Can be Improved by Accounting for Serial Correlation Through an Autoregressive Model

Qian Zhang1 and Robert M. Hirsch2

1University of Maryland Center for Environmental Science/U.S. Environmental Protection Agency Chesapeake Bay Program, Annapolis, MD, USA, 2U.S. Geological Survey, Reston, VA, USA

Abstract Accurate quantification of riverine water-quality concentration and flux is challenging because monitoring programs typically collect concentration data at lower frequencies than discharge data. Statistical methods are often used to estimate concentration and flux on days without observations. One recently developed approach is the Weighted Regressions on Time, Discharge, and Season (WRTDS), which has been shown to provide among the most accurate estimates compared to other common methods. The main objective of this work was to improve WRTDS estimation by accounting for the autocorrelation structure of model residuals using the first-order autoregressive model (AR1). This modified approach, called WRTDS-Kalman Filter (WRTDS-K), was compared with WRTDS for six constituents including nitrate-nitrogen (NOx), total phosphorus, total Kjeldahl nitrogen, soluble reactive phosphorus, suspended sediment, and chloride. Near-daily concentration records at nine sites were used to generate subsets through Monte Carlo sampling for five different sampling scenarios. Results show that WRTDS-K provided generally better daily estimates of concentration and flux than WRTDS under these sampling scenarios for all constituents, especially NOx. The degree of improvement is strongly affected by the underlying sampling scenario, with WRTDS-K gaining more advantage when more samples are available, and hence more residuals can be exploited. The performance of WRTDS-K depends on the AR1 coefficient (ρ) and that relationship varies with constituents and sampling scenarios. These results provided recommendations on the optimal ρ for each constituent and sampling scenario. Overall, WRTDS-K has the potential for broad applications to monitoring records elsewhere, as demonstrated by a pilot application to Chesapeake Bay tributaries.

1. Introduction

Surface water-quality degradation has negative effects on the health of ecosystems and abundance of biological living resources. A fundamental challenge to the effective management of surface water quality is the accurate quantification of the flux of materials (e.g., sediment and nutrients) that are transported through rivers and streams. Constituent flux, representing the mass of materials passing a particular location in a given period of time, is critical for the establishment of ecosystem restoration targets (e.g., maximum daily loads; U.S. Environmental Protection Agency, 2010), calibration of watershed models (e.g., SPARROW, HSPF; Ator et al., 2011; Ator et al., 2019; Linker et al., 2013; Shenk & Linker, 2013), and evaluation of water-quality trends and drivers (Bowes et al., 2008; Green et al., 2014; Hirsch, 2012; Zhang et al., 2015).

Accurate quantification of constituent concentration and flux is challenging because cost considerations usually dictate that water-quality observations are typically only made at frequencies such as weekly or monthly even though discharge data are available at a daily or more frequent time step. For example, sites in the Chesapeake Bay watershed (USA) are sampled at least 20 times per year, including at least 12 regular monthly samples and at least eight stormflow samples (Chanat et al., 2016). In comparison, concentration data at many other locations are even more sparse and often do not contain targeted stormflow samples.

To estimate complete records of daily concentration and daily flux (the product of daily concentration and daily discharge), a number of statistical approaches have been developed, including but not limited to Cohn et al. (1989), Cohn et al. (1992), Johnes (2007), and Park and Engel (2015). These approaches are based on statistical relations between measured concentrations (the response variable) and a set of explanatory variables that typically include time, season, and discharge associated with the date of sample collection. In virtually all these models, a single regression model developed with all available data in the period of record is used to estimate concentrations and fluxes for unsampled days.
These statistical approaches rely on assumptions of homoscedasticity of model errors and fixed relations between concentration and each covariate, which can be frequently violated in real water-quality data—see examples in Hirsch et al. (2010) and Zhang et al. (2016). To relax such model assumptions, Hirsch et al. (2010) developed Weighted Regressions on Time, Discharge, and Season (WRTDS). Like many of its predecessors, WRTDS uses time, discharge, and season as explanatory variables. However, WRTDS differs in that it functionally develops a separate model for each day of the observed record by re-evaluating the dependencies of concentration on time, discharge, and season using samples most relevant to the day of estimation (Hirsch et al., 2010; Hirsch & De Cicco, 2015). The method also removes the constraint of homoscedastic residuals that is typical of regression-based models. Consequently, WRTDS can better represent the temporally-varying seasonal and discharge-related water-quality patterns and thus produce more accurate estimates (Chanat et al., 2016; Hirsch, 2014; Lee et al., 2016, 2019; Moyer et al., 2012). For example, Chanat et al. (2016) compared WRTDS and the seven-parameter LOADEST model (Cohn, 2005; Cohn et al., 1989; Cohn et al., 1992) for 80 sites in the Chesapeake Bay watershed and observed generally greater explanatory power by WRTDS. Since its publication, WRTDS has been adopted in a range of surface water-quality studies in the United States and elsewhere, including nitrogen (Ator et al., 2019; Chanat & Yang, 2018; Sprague et al., 2011; Strickling & Obenour, 2018; Van Meter et al., 2017), phosphorus (Fanelli et al., 2019; Rankinen et al., 2016), sediment (Hirsch, 2012; Zhang et al., 2016), chloride (Corsi et al., 2015; Stets et al., 2018), organic carbon (Zhang & Blomquist, 2018), organic matter (Stackpoole et al., 2017), and major ions (Bird et al., 2018).

Although WRTDS is more likely to compute accurate load estimates than other commonly used regression methods, there are potential opportunities to further improve its model performance. These improvements can be realized by investigating model errors that are not accounted for by the three model covariates (i.e., time, discharge, and season). In this regard, one potential approach is to consider new explanatory variables to further account for variability in sampled concentrations. For example, it is often reported that riverine transport can be affected by watershed antecedent conditions due to the buildup of momentum and activation of new pathways (Davis et al., 2014; Gray et al., 2015; Macrae et al., 2010; Warner et al., 2009). To quantify antecedent conditions, Vecchia (2003) has developed the “flow anomaly” concept, which can be computed based on daily discharge records that are widely available. In the context of WRTDS, this concept was first considered by Murphy et al. (2014) to focus on nitrate transport. Expanding upon the effort of Murphy et al. (2014), Zhang and Ball (2017) developed and tested a set of modified WRTDS models with flow-anomaly variables and reported generally improved estimates of constituent concentrations and fluxes by the modified models.

While it can be useful to consider new explanatory variables within WRTDS, another potential opportunity for improvement is to investigate the autocorrelation structure of the WRTDS model residuals and exploit that information to adjust estimates. In other words, WRTDS estimates could be adjusted through an empirical residual correction approach to better approximate observations. Such a concept has been previously tested in the context of LOADEST (Appling et al., 2015; Aulenbach, 2013; Aulenbach & Hooper, 2006; Verma et al., 2012). These methods are collectively referred to as the “composite method” and have been found to provide more accurate estimates through the empirical residual correction. However, these studies were limited to LOADEST and limited to simple autocorrelation structures for concentration residuals. Specifically, Aulenbach and Hooper (2006) used a piece-wise linear interpolation of residuals, whereas Verma et al. (2012) used rectangular- and triangular-shaped patterns to capture the autocorrelation structure. Similarly, Schwarz et al. (2006) developed the FLUXMASTER method, which also has the capacity of adjusting estimates based on model residuals. However, no time series model has been invoked in this context.

Thus, the overall objective of this work is to test the hypothesis that WRTDS estimation can be improved by accounting for the autocorrelation structure of model residuals. Previous studies of time-series residuals often exhibit strong serial correlation at time scales of days and up to many months (Gray, 2018; Hirsch & Slack, 1984; Kirchner & Neal, 2013; Zhang et al., 2018). Serial correlation is generally related to specific characteristics of individual hydrologic events (high flows or drought periods) that are not captured by the statistical model being used. These characteristics may include differences in antecedent conditions (see Murphy et al., 2014; Vecchia, 2003) or increased (or decreased) hydrologic contributions from subwatersheds that differ in terms of sources or dominant transport processes. An extreme example of this is the Missouri River flood of 2011, as documented by Kalkhoff (2013).
To limit the scope of the work, we focused on the first-order autoregressive (AR1) model for capturing residual autocorrelation. We call this modified approach “WRTDS-Kalman Filter (WRTDS-K)” because it uses the residual information on the sampled days to produce adjusted estimates for the unsampled days. We compared the performance between WRTDS-K and WRTDS for nitrate-plus-nitrite (NOx), total phosphorus (TP), total Kjeldahl nitrogen (TKN), soluble reactive phosphorus (SRP), suspended sediment (SS), and chloride (Cl). Method performance was compared using near-daily data at nine monitoring sites. For each site and constituent, concentration subsets were generated through Monte Carlo subsampling for five sampling scenarios and used for model implementation and evaluation. Specific goals of this work included the following:

1. To test the hypothesis that WRTDS-K can improve the estimation of concentration and flux of common water-quality constituents, particularly NOx for which discharge is often a rather poor predictor.
2. To search for the optimal value(s) of the lag-1 coefficient (ρ) for different water-quality constituents and sampling scenarios.
3. To apply WRTDS-K (with appropriate ρ) to major tributaries to Chesapeake Bay and compare the WRTDS-K estimates with the regular WRTDS estimates.

The results from this research will inform the design and management of riverine water-quality monitoring networks as well as the analysis of water-quality data for concentration and flux estimation. Thus, this work is expected to have broad implications toward traditional water-quality monitoring sites that are distributed across the world.

It should be noted that WRTDS-K is only intended for use when the goal is to provide the best possible estimates on the actual daily values of concentration (or flux) or their mean values aggregated over periods such as months, seasons, or years. The original WRTDS method also produces estimates of the flow-normalized concentration and flux values, which are used in evaluating long-term trends in water quality (Hirsch et al., 2010; Hirsch & De Cicco, 2015). Such trends will continue to be computed using the original model.

2. Methods
2.1. WRTDS

The implementation of the typical WRTDS model is illustrated in Figure 1a. The core equation of the model can be mathematically described as follows (Hirsch et al., 2010):

\[
\ln(C_i) = \beta_{0,i} + \beta_{1,i} \ln(Q_i) + \beta_{2,i} T_i + \beta_{3,i} \sin(2\pi T_i) + \beta_{4,i} \cos(2\pi T_i) + \sigma \varepsilon_i
\]

where

- \(C_i\) is the constituent concentration on day \(i\), in mg/L;
- \(Q_i\) is the daily mean discharge on day \(i\), in m\(^3\)/s;
- \(T_i\) is the time expressed as decimal year;
- \(\beta_{0,i}, \beta_{1,i}, \beta_{2,i}, \beta_{3,i},\) and \(\beta_{4,i}\) are the fitted coefficients that vary smoothly over

Figure 1. Flow-chart illustration of a the regular WRTDS model and b the WRTDS-K model. WRTDS-K differs from WRTDS in terms of model estimation (mainly the adjustment of standard residuals based on actual observations for the sampled days and AR1 model interpolation for the unsampled days) and model output.
the model domain; \( \sigma_i \) is the fitted value of the conditional standard deviation of the error of the model; and \( z_i \) is the standardized model residual on day \( i \) (standard deviation = 1).

The \( \ln(C) - \ln(Q) \) relationship is adopted for three reasons: (1) it provides a generally better fit than a linear relationship, (2) it eliminates the possibility of obtaining negative estimates for concentration, and (3) the residuals from this model are more nearly normal (Hirsch & De Cicco, 2015).

Functionally, WRTDS develops one regression model for each day of the record to estimate concentration and flux. With the known time, discharge, and season of the estimation day, WRTDS prescreens the entire concentration record and selects samples that are sufficiently “close” to the estimation day with respect to time, discharge, and season. The selected samples are used to build a weighted regression model using equation (1) and a unique set of model coefficients is obtained. These coefficients are used to estimate \( \ln(C) \) on the estimation day by substituting known values of time and discharge, which is then transformed to concentration in real space using a bias correction factor (BCF; equation (3))

\[
E[C_i] = \exp\{\beta_{0i} + \beta_1 \ln(Q_i) + \beta_2 T_i + \beta_3 \sin(2\pi T_i) + \beta_4 \cos(2\pi T_i) + BCF_i\}
\]

\[
BCF_i = \frac{\sigma_i^2}{2}
\]

The BCF arises because we seek to estimate the conditional mean of concentration, but the model is designed to estimate the conditional mean of the natural log of concentration. This BCF is approximately correct when the errors in log space are normal and sigma is relatively small, and the sample size is large (e.g., >50).

To expedite the estimation process, the above steps are conducted on a grid network formed by \( t \) and \( \ln(Q) \). For the \( t \) axis (\( x \) axis), grid values are spaced 1/16th of a year apart from the beginning year to the end year of the record. For the \( \ln(Q) \) axis (\( y \) axis), 14 grid values are spaced with equal distance for the discharge range from 5% below the minimum discharge to 5% above the maximum discharge in the record. For each grid point, WRTDS develops a separate weighted regression model, which results in an estimated concentration “surface” as a function of \( t \) and \( \ln(Q) \). Daily concentration is then estimated using a bilinear interpolation of the surface, which is multiplied by daily discharge to compute daily flux. For more details, see Hirsch and De Cicco (2015).

### 2.2. WRTDS-K

WRTDS-K is a variation of the regular WRTDS method for estimating concentration as a function of time, discharge, and season (Figure 1b). First, the method develops the regular WRTDS model using the available concentration record. Second, the method calculates the residuals in the logarithm space for days with observed concentrations, that is,

\[
r_i = \ln(C_i) - (\beta_{0i} + \beta_1 \ln(Q_i) + \beta_2 T_i + \beta_3 \sin(2\pi T_i) + \beta_4 \cos(2\pi T_i))
\]

These residuals are the error in the model, expressed as the observed \( \ln(C) \) minus predicted \( \ln(C) \). Third, the method standardizes these model residuals by dividing them by the standard deviation appropriate to that day (see equation (1)), resulting in the standardized residuals, \( z_i \):

\[
z_i = \frac{r_i}{\sigma_i}
\]

In this work, WRTDS-K assumes that the standardized residuals have a serial correlation structure of AR1 with a lag-1 correlation coefficient (\( \rho \)) that falls in the range of (0, 1). It is recognized that AR1 may not be the correlation structure that best represents the behavior of the standardized residuals, but simulation of more complex correlation structures is generally not feasible given the irregular and sparse sampling frequencies of most water-quality data sets. Previous study of selected daily water-quality records have shown that an AR(1) model may understate the correlation at lag times such as a week or a month; however, the authors believe that the AR(1) model is the most feasible to incorporate the residual information and still produce useful results.
4. This time series of \( z \) values is transformed to a set of concentration values:

\[
\text{Equation (3)}
\]

Based on the estimated values of \( \text{Equation (3)} \), is not used in WRTDS which is equivalent to the composite method (Aulenbach & Hooper, 2006).

50 replicates in the WRTDS were used for all our WRTDS based on 1,000 replicates the generated values are solely based on \( ek \).

The above process is repeated \( M \) times and the expected value of concentration for each day \( i \) is the mean of the \( M \) replicates of \( C_i \), which is the expected value of concentration for day \( i, \overline{C}_i \). Note that the BCF, defined in Equation (3), is not used in WRTDS-K. It is used in the original formulation of WRTDS because it estimates a mean value based on the conditional moments of the logarithms. In WRTDS-K, in equation (7), the estimates are for a single value rather than a mean value. The estimate of the mean is computed only after retransforming each of the individual replicate values for the day. To identify an appropriate value for \( M \), we explored the sensitivity of \( \overline{C}_i \) to \( M \) and found that \( \overline{C}_i \) values based on 50 replicates were very close to values based on 1,000 replicates—see a case study in Figure S1 in the supporting information. Thus, 50 replicates were used for all our WRTDS-K model runs in this study. There are data sets elsewhere for which using 50 replicates in the WRTDS-K calculations may not be sufficient to achieve stable results. It can be useful to run the analysis a second time with a different random number seed and determine if the new set of results (when aggregated to monthly or annual totals) differs substantially from the first set of results. If that is the case, then one could increase the number of replicates to several hundred. The computational burden of using a larger number of replicates is relatively small.

For **sampled days**, the estimate of concentration is equal to the measured value. By contrast, for the regular WRTDS model the estimate of concentration obtained using equation (2) is the unbiased estimate of concentration. Thus, WRTDS-K will always be more accurate than WRTDS for the sampled days. For **unsampled days**, the estimate of concentration by WRTDS-K is also expected to be different from that by WRTDS due to the step of empirical residual correction. In general, when the sampling is sparse, such as a typical gap of more than 60 days, WRTDS-K should produce estimates that are similar to those determined in the regular WRTDS model. When the typical gaps are small, such as seven days or less, WRTDS-K estimates can be quite different from WRTDS estimates because there are many measured data values that WRTDS-K can use and the serial dependence of the data at short lags (e.g., one to seven days) have increased influence on the estimates.
Another way to conceptualize the operation of WRTDS-K is to consider that on any given day there are three sources of information available to make our daily estimate. One source is the most recent residual from before the day being estimated. The second is the subsequent residual after the day being estimated. The third is zero, which represents the regression model (with no error), which provides the best available representation of water-quality transport processes. The task then is to come up with a single estimate that is the weighted sum of these three bits of information. The weights of the adjacent samples are rather high when they are very close to the sampled days and low when there is a long period before or after them. The approach presented is viewed as a reasonable and objective way to set those weights.

3. Sites, Data, and Experiments

3.1. Testing Sites and Data

The development and testing of the WRTDS-K approach was done using near-daily water-quality data for several tributaries to Lake Erie and the Ohio River, which are available through the National Center for Water Quality Research at Heidelberg University, OH, USA (National Center for Water Quality Research, 2015). The data sets cover nine monitoring sites that had at least 15 years of concentration samples, located within the Maumee River, Scioto River, Great Miami River, Sandusky River, River Raisin, Cuyahoga River, Grand River, Honey Creek, and Rock Creek (Figure 2). These sites cover a range of drainage area (34.6–6,330 mi²), agricultural land (30.4%–89.9%), wooded land (7.3%–50.1%), and urban land (0.6%–9.6%). Consequently, the results from this research are expected to be not limited to a single watershed or watershed type within this geographic setting (Table S1 in the supporting information).

Figure 2. The Lake Erie and Ohio River tributaries that have been monitored by the National Center for water quality research at Heidelberg University, OH, USA. Details of the selected sites and available water-quality data are summarized in tables S1 and S2, respectively.
This region has the following general characteristics: (1) it is located in the north temperate climatic zone, (2) its physiography is dominated by the Central Lowlands of the Interior Plains Division, and (3) it is underlain with thousands of feet of sedimentary rocks formed during the Paleozoic Era (Schiefer, 2002). For characteristics of each tributary, see Schiefer (2002).

For each site and constituent, the original concentration data have between 256 and 349 sampled days per year on average (Table S2). The median of sampled concentrations was used for any day in which multiple concentration values were recorded. Daily river discharge data at each site are available for these sites through the U.S. Geological Survey (USGS) National Water Information System database (U.S. Geological Survey, 2018). Concentration and discharge previously downloaded and processed by Zhang and Ball (2017) were used here.

3.2. Data Subsampling

To test the performance of WRTDS-K, we selected five sampling scenarios and performed Monte Carlo subsampling of the original concentration data to produce concentration subsets (30 subsets per sampling scenario) to serve as the “training data” for subsequent model run. These scenarios are as follows:

S1: One concentration sample randomly selected for each calendar month for which concentration samples were available (i.e., 12 concentration samples per year),

S2: Flow-stratified sampling with 12 regular samples and 8 storm samples per year (i.e., 20 samples per year). Within each specific year of the record, samples with daily discharge exceeding the 90th percentile of daily discharge distribution of the year were marked as “storm sample”; all other samples were considered as “regular samples.” Because a similar sampling scenario has been adopted for many sites in the Chesapeake Bay watershed—see details in Chanat et al. (2016), evaluation of model performance for this strategy can provide important insights toward the Chesapeake Bay watershed monitoring network and similar programs elsewhere,

S3: Twelve random monthly samples (i.e., S1 above) plus additional eight random samples per year (i.e., 20 samples per year),

S4: Two samples randomly selected for each calendar month for which concentration samples were available (i.e., 24 samples per year), and

S5: One sample randomly selected for each week for which concentration samples were available (i.e., 52 samples per year).

For illustration, an example of NO₃ in the Maumee River at Waterville, OH is provided in Figure 3, which shows the original concentration data and the five types of subsamples.

3.3. Model Implementation and Evaluation

With each concentration subset, WRTDS and WRTDS-K models were run to estimate daily concentration and flux. WRTDS was implemented in R using the package “EGRET” (Hirsch & De Cicco, 2015). WRTDS-K was implemented in R with additional codes that implement equations (4)–(7). To test the appropriateness of the AR1 serial correlation structure, a set of ρ values were used, namely, 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.825, 0.85, 0.875, 0.9, 0.925, 0.95, 0.975, and 0.99.

To quantify model performance, concentration and flux values in the original data set that were not selected in the subsamples (section 3.2) were used for model validation. Based on these data, the modified Nash-Sutcliffe efficiency (MNSE) was calculated:

$$MNSE = 1 - \frac{\sum_{i=1}^{n} |Y_{i}^{\text{obs}} - Y_{i}^{\text{ref}}|}{\sum_{i=1}^{n} |Y_{i}^{\text{obs}} - Y_{i}^{\text{mean}}|}$$

where $Y_{i}^{\text{obs}}$ and $Y_{i}^{\text{ref}}$ are the observed and modeled values for the $i$th day, respectively. $Y_{i}^{\text{mean}}$ is the average of observed data in the validation data, and $n$ is the total number of observations in the validation data. MNSE quantifies the absolute difference between the estimated and observed values normalized by the sum of the absolute deviations of the observations. MNSE is less sensitive to large errors than the standard NSE, because it uses absolute difference whereas the latter uses squared difference (Krause et al., 2005; Moriasi et al., 2007). In general, a larger MNSE means a better model and a model with MNSE = 1 implying that the estimates match perfectly with observations.
We followed the approach of Zhang and Ball (2017) to repeat the concentration subsampling experiment 30 times for each site-constituent combination. For each subset, we quantified the MNSE for WRTDS and WRTDS-K (with varying ρ) for each of the 30 replicate runs and the median MNSE was recorded and used for subsequent model comparison. This procedure helps reduce the required computational time, while providing a reasonable number of replicates to describe overall uncertainties in the model performance.

The above analysis resulted in MNSE values for WRTDS and WRTDS-K (with varying ρ) for each sampling scenario for each of the 54 site-constituent combinations (i.e., 9 sites × 6 constituents). To more directly compare the results, the ratio between the MNSE values of the two methods (R) was calculated for each case of site, constituent, and sampling scenario:

\[ R(\rho) = \frac{\text{MNSE}_{\text{WRTDS-K}}(\rho)}{\text{MNSE}_{\text{WRTDS}}} \]  \hspace{0.5cm} (9)

Note that R values greater than 1 indicate that the WRTDS estimation is improved by WRTDS-K. These R values were used to determine the optimal ρ value(s) for the six constituents under different sampling scenarios.

While our comparison focused on MNSE, we used percent bias (PB) as an auxiliary metric for evaluating model performance for the entire period of record:

\[ PB = \frac{\sum_{i=1}^{n} (Y_{i}^{\text{obs}} - Y_{i}^{\text{est}})}{\sum_{i=1}^{n} Y_{i}^{\text{obs}}} \]  \hspace{0.5cm} (10)

PB could be negative, positive, or zero; a value of zero means that the model is unbiased.

Figure 3. Example time series of (a) the original NOx concentration data (black open circles) and (b–f) the five types of subsamples in Maumee River at Waterville, OH in the calendar year of 1998. For (b–f), the original data are shown as black dots and subsamples are shown as solid red circles. Inset text shows the number of (a) the original data and (b–f) the subsamples in the year of 1998.
3.4. Ancillary Analysis of Continuous-Monitoring Data

To further characterize optimal \( \rho \) values, we conducted an independent analysis on continuous-monitoring data at three additional monitoring stations. These stations are North Raccoon River near Sac City, IA (USGS ID: 05482300); Potomac River near Washington, D.C. (Little Falls; 01646500); and Mississippi River at Baton Rouge, LA (07374000), and they vary in 3 orders of magnitude in terms of drainage area. This preliminary analysis was limited to NO\(_x\) due to data availability. For each site, 15-min measurements of NO\(_x\) were downloaded from the USGS National Water Information System database (U.S. Geological Survey, 2018). These data were processed as follows:

1. Subsample the 15-min record to a daily record by selecting the values recorded at the time 12:00 of each day,
2. Run the regular WRTDS model on the daily record to obtain daily concentration estimates,
3. Calculate the standardized residuals (\( z_i \)) using equations (4) and (5),
4. Extract the longest contiguous sequence (i.e., gap free) in \( z_i \),
5. Estimate \( \rho \) and the autocorrelation function (ACF) for the extracted sequence using the R function ar (order.max = 1, method = “yule-walker”) (R Core Team, 2019).

These estimates of \( \rho \) were then compared with the optimal \( \rho \) identified using Monte Carlo analysis of the Heidelberg data sets (section 3.3).

3.5. Model Application to Chesapeake Tributaries

A pilot application of the WRTDS-K approach was implemented with optimal \( \rho \) values on water-quality data collected from major tributaries draining to Chesapeake Bay, the largest estuary in the United States. We focused on the USGS River Input Monitoring Program stations that collectively account for a majority of nutrient and sediment inputs from the nontidal Chesapeake watershed (Table S3 and Figure S2). For these sites, sampling scenario S2 is considered to best approximate actual sampling procedures (Chanat et al., 2016). Water-quality and discharge data at these sites were available through Moyer et al. (2017). Cl was excluded because it had not been a focus of Chesapeake watershed monitoring. For each river and constituent, flux estimates from WRTDS and WRTDS-K were compared on daily, monthly, and annual scales. Mean absolute difference (MAD) was used to quantify the level of difference between the two sets of estimates:

\[
MAD = \frac{\sum_{i=1}^{n} |K_{est}^i - W_{est}^i|}{n}
\]  

where \( W_{est}^i \) and \( K_{est}^i \) represent the WRTDS and WRTDS-K flux estimate for time index i, respectively. This metric was computed for annual (where \( n = \) number of years \( \times 365 \)), monthly (\( n = \) number of years \( \times 12 \)), and annual scales (\( n = \) number of years), respectively.

4. Results and Discussion

4.1. WRTDS-K Versus WRTDS: The NO\(_x\) Example

To guide readers through our results, we start with a preliminary comparison of the two models by focusing on NO\(_x\) and a preselected \( \rho \) (i.e., 0.95) for WRTDS-K (Figure 4). For NO\(_x\) concentration (Figure 4a), two key patterns emerged. First, WRTDS-K performed better than WRTDS for almost all sites and sampling scenarios. Second, the degree of improvement achieved by WRTDS-K, represented by the vertical distance of the data points from the 1:1 line, varied considerably among the different sampling scenarios. Among the five sampling scenarios, S1 generally had the least amount of improvement, whereas S5 showed the greatest improvement. Since both scenarios are based on random sampling, this difference is attributable to the different frequencies of concentration samples—that is, 12 versus 52 per year for S1 and S5, respectively. As expected, this result shows that WRTDS-K becomes more accurate relative to WRTDS when more concentration samples are available, and thus, more residuals can be utilized in the AR1 interpolation procedure (equation (6)). For NO\(_x\) flux (Figure 4b), it is also evident that WRTDS-K performed generally better than WRTDS, but the degree of improvement was lower than in the case of NO\(_x\) concentration. In addition, the degree of improvement by WRTDS-K was less variable among the sampling scenarios. Overall, this preliminary comparison provides three key messages. First, WRTDS-K can provide improved estimates of NO\(_x\)
compared to WRTDS for almost all sites under the five different sampling scenarios. Second, the degree of improvement in concentration estimation is strongly affected by the underlying sampling scenario, with most improvement achieved by S5. Third, the degree of improvement by WRTDS-K is more evident for concentration than flux.

4.2. Sensitivity of WRTDS-K Performance to $\rho$

Here we consider results from all sites and constituents to examine the sensitivity of the WRTDS-K performance to the choice of $\rho$. For better communication of the results, the median of the computed $R$ values (equation (9)) of all sites was quantified and plotted as a function of $\rho$. Recall that $R$ values greater than 1 indicate that WRTDS-K estimates have a higher MNSE than WRTDS estimates (equation (9)).

Figure 5 shows the site-integrated results for NOx, which reinforce the messages presented above. Generally, $R$ values are greater than 1.0, meaning that WRTDS-K produces more accurate estimates than WRTDS for almost the entire range of $\rho$. As expected, among all sampling scenarios, WRTDS-K does not show improvement over WRTDS when $\rho$ is zero, since the estimation is not informed by residuals on sampled days. Second, as illustrated in Figure 4, the greatest improvement by WRTDS-K is achieved with sampling scenario S5 and the least improvement with S1. Also, as illustrated previously, the degree of improvement by WRTDS-K is more evident for concentration than flux. The relative improvement by S2 with respect to NOx flux estimation is even greater than that by S4, stressing the value of stormflow sampling in the context of residual utilization by WRTDS-K. Finally, Figure 5 shows that $R$ gradually increases with $\rho$ and the optimal $\rho$ is roughly in the range of 0.9–0.99 for concentration estimates and 0.95–0.99 for flux estimates. This pattern holds true for all five sampling scenarios and shows that the performance of WRTDS-K relative to WRTDS is sensitive to the choice of $\rho$ for NOx.

Similar plots are displayed for TP (Figure S3), TKN (Figure S4), SRP (Figure S5), SS (Figure S6), and Cl (Figure S7) to examine the sensitivity of WRTDS-K performance to $\rho$. These results demonstrate similar results as presented for NOx and show that the performance of WRTDS-K relative to WRTDS varies with $\rho$. Importantly, the optimal value (or range) of $\rho$ varies among sampling scenarios and among constituents, which should be considered when developing recommendations on the optimal $\rho$ for WRTDS-K. The patterns of $R$ versus $\rho$ for these constituents are generally different from the pattern for NOx (Figure 5). For example, TP results showed that $R$ gradually increases with $\rho$ between 0 and 0.875 (Figure S3). Through visual inspection of these plots, the optimal range of $\rho$ was identified to be roughly 0.8–0.875, 0.8–0.9, 0.8–0.925, 0.8–0.925, and 0.8–0.95 for TP, TKN, SRP, SS, and Cl, respectively.

Overall, these results demonstrate that (1) WRTDS-K can lead to improved estimation for a large range of $\rho$ through the utilization of model residuals on sampled days, (2) the most improvement with WRTDS-K is generally achieved with $\rho$ values greater than ~0.8, and (3) the optimal $\rho$ varies among constituents and among sampling scenarios.
4.3. Recommended $\rho$ for Different Constituents and Sampling Scenarios

While the above site-integrated results provide insights on the optimal $\rho$, they do not speak to the differences among the nine study sites. In this regard, we developed a simple and objective algorithm to search for the optimal $\rho$ for concentration estimation. For each combination of site, constituent, and sampling scenario, we ranked the $R$ values for the nine sites and selected the $\rho$ values corresponding to the three largest $R$ values. We then counted the occurrence of the selected $\rho$ values for each constituent and each sampling scenario and plotted the counts as functions of $\rho$ and sampling scenario for each constituent (Figure 6).

For sampling scenario S1, the optimal values of $\rho$ appear to be $0.925$–$0.99$, $0.85$–$0.925$, $0.825$–$0.9$, $0.85$–$0.9$, and $0.875$–$0.9$ for NO$_x$, TP, TKN, SRP, SS, and Cl, respectively. These optimal values are generally in the vicinity of $0.9$. For sampling scenario S2, the optimal values of $\rho$ are $0.925$–$0.99$, $0.875$–$0.925$, $0.85$–$0.9$, $0.85$–$0.9$, $0.875$–$0.925$, and $0.85$–$0.95$ for NO$_x$, TP, TKN, SRP, SS, and Cl, respectively, which are similar to those for sampling scenario S1. These optimal values—and those identified for S3, S4, and S5—are summarized in Table 1, which provide useful guidelines for applying WRTDS-K to other monitoring data sets.

The recommended values of $\rho$ (Table 1) are further inspected against the $R$ versus $\rho$ plot of each individual site-constituent case (plots not shown). In most cases, the optimal range of $\rho$ is generally consistent with the range listed in Table 1. Also, in most cases, $R$ values are greater than 1.0, indicating improved performance by WRTDS-K than WRTDS. Moreover, the shape of the $R$ versus $\rho$ plots are also comparable among the nine different sites in most cases—that is, $R$ starts at a value of ~1.0 at $\rho = 0$, increases steadily with $\rho$, and reaches...

Figure 5. The median ratio of MNSE for WRTDS-K to MNSE for WRTDS for the nine Lake Erie/Ohio River sites for (a) daily NO$_x$ concentration and (b) daily NO$_x$ flux, shown as a function of $\rho$ in AR1. The legend in (a) lists the five sampling scenarios—See section 3.2.

Table 1. Recommended Values of $\rho$ for Each Constituent and Sampling Scenario

| Constituent | S1 | S2 | S3 | S4 | S5 |
|-------------|----|----|----|----|----|
| NO$_x$      | 0.925–0.99 | 0.85–0.925 | 0.825–0.9 | 0.85–0.9 | 0.875–0.9 |
| TP          | 0.925–0.99 | 0.875–0.925 | 0.85–0.9 | 0.85–0.9 | 0.875–0.925 |
| TKN         | 0.925–0.99 | 0.875–0.925 | 0.85–0.9 | 0.85–0.9 | 0.875–0.925 |
| SRP         | 0.925–0.99 | 0.875–0.925 | 0.85–0.9 | 0.85–0.9 | 0.875–0.925 |
| SS          | 0.925–0.99 | 0.875–0.925 | 0.85–0.9 | 0.85–0.9 | 0.875–0.925 |
| Cl          | 0.925–0.99 | 0.875–0.925 | 0.85–0.9 | 0.85–0.9 | 0.875–0.925 |

10.1029/2019WR025338

Zhang and Hirsch 9715
a plateau at \( \rho = 0.7 \) or 0.8, which may continue through \( \rho = 0.99 \) or decrease slightly in the range of \( \rho = 0.9 \)–0.99, depending on the constituents and sampling scenarios. Broadly speaking, the recommended values of \( \rho \) (Table 1) should be most applicable to monitoring data collected in geological and hydrological settings that are similar to the Lake Erie and Ohio River region. Caution should be exercised if these recommended values are applied to monitoring records obtained in settings that are substantially different from this region. Unfortunately, long records of near-daily water quality

Table 1

| Constituent | S1       | S2       | S3       | S4       | S5       |
|-------------|----------|----------|----------|----------|----------|
| NO\(_x\)    | 0.925–0.99 | 0.925–0.99 | 0.925–0.99 | 0.9–0.975 | 0.95–0.99 |
| TP          | 0.85–0.925 | 0.875–0.925 | 0.8–0.9 | 0.8–0.9 | 0.85–0.9 |
| TKN         | 0.825–0.9 | 0.85–0.9 | 0.8–0.925 | 0.8–0.925 | 0.7–0.875 |
| SRP         | 0.85–0.9 | 0.85–0.9 | 0.825–0.9 | 0.8–0.875 | 0.8–0.85, 0.925 |
| SS          | 0.9 | 0.875–0.925 | 0.875–0.9 | 0.9–0.925 | 0.85–0.9875, 0.925 |
| Cl          | 0.85–0.925 | 0.85–0.95 | 0.85–0.925 | 0.85–0.925 | 0.85–0.925 |

Figure 6. Count of \( \rho \) values that resulted in the greatest improvement in daily concentration estimation by WRTDS-K, shown as functions of \( \rho \) and sampling scenarios. Note that the counts cannot exceed 9 (i.e., nine sites).
data are uncommon outside of this region. It would be useful to conduct additional studies of the optimal $\rho$ values in cases where these are available. High-frequency sensors of NO$_x$, SRP, and turbidity do exist in a growing number of locations and may prove suitable for such studies to further refine the selection of optimal $\rho$ values.

To further support the validity of these results, we estimated the AR1 coefficient for continuously monitored NO$_x$ records at three selected sites. Figure 7 shows the longest contiguous sequence of standardized residuals for these sites, as obtained using the procedure described in section 3.4. These results show that standardized residuals behave differently among the sites for the selected time periods. For the North Raccoon River, which has the smallest drainage area among the three sites, the standardized residuals shift from generally positive values to generally negative values at about the 200th day of the 439-day period (Figure 7a) and the ACF exhibits a long memory (i.e., high correlation even at a lag of 30; Figure 7b). For the Mississippi River, which has the largest drainage area among the three sites, the standardized residuals remain generally positive in this 266-day period (Figure 7e) and the ACF switches from positive correlations at small lags to
negative correlations at large lags (Figure 7f). For these two sites, $\rho$ is estimated to be 0.93 and 0.956, respectively, which are consistent with the recommended range for NO$_x$ that were derived from the Monte Carlo analysis (Table 1 and Figure 6a). Standardized residuals from the Potomac River near Washington D.C. site have positive values on most days in the sequence except for several events (Figure 7c). The associated ACF exhibits a long memory (>30 lags) and the estimated $\rho$ is 0.868 (Figure 7d). On the one hand, this analysis demonstrates that the recommended range of $\rho$ for NO$_x$ (Table 1) is generally supported with these independent, continuous-monitoring data sets. On the other hand, this analysis suggests that the AR1 model may not always be the most appropriate assumption for capturing the ACF of the standardized residuals, since the correlation can persist at lags much longer than one day. As noted above, simulation of more complex correlation structures is generally not feasible given the irregular and sparse sampling frequencies of most water-quality data sets. Analysis of the correlation structure of model residuals of water-quality sensor data is an important topic for future research, which is only beginning to be possible as long-term sensor records become available.

4.4. Summary of Improvement by WRTDS-K (With Optimal $\rho$)

With the optimal $\rho$ identified for the different sampling scenarios, the improvement by WRTDS-K relative to WRTDS can be quantified accordingly. Figure 8 compares the MNSE of WRTDS and WRTDS-K (with optimal $\rho$) for concentration for the six constituents. The optimal $\rho$ value used here corresponds to the largest count in Figure 6. As expected, WRTDS-K generally shows a higher MNSE than WRTDS. (For comparison on flux estimates, see Figure S8.)

Table 2 summarizes the percent improvement in MNSE by WRTDS-K relative to WRTDS. For simplicity, only the median values of the nine study sites are presented, which are 56–119%, 13–36%, 23–61%, 16–45%, 10–20%, and 21–59% for NO$_x$, TP, TKN, SRP, SS, and Cl, respectively (Table 2). The lower and upper ends of these values correspond to S1 and S5, respectively. Comparing the six constituents, the smallest improvement by WRTDS-K is associated with SS, whereas the greatest improvement is associated with NO$_x$. Again, these results demonstrate that WRTDS can be further improved by WRTDS-K through the exploitation of the sample-associated residuals, especially for monitoring records with high sampling frequencies.

The two models were briefly compared using the percent bias (PB) metric (equation (10)) using optimal $\rho$ values identified above. The results from this comparison are summarized in Figures S9 and S10. For daily concentration estimates, WRTDS-K is less biased than WRTDS in almost all cases (i.e., five sampling scenarios $\times$ nine study sites) for TP, TKN, and especially SRP (Figure S9b–S9d) and a majority of cases for NO$_x$ (Figure S9a). For SS, WRTDS-K appears to be less biased than WRTDS in about one half of the cases but more biased in the other half (Figure S9e). For Cl, both methods are nearly unbiased in all cases (Figure S9f). For daily flux estimates, the two approaches appear to have similar yet small biases in all cases for NO$_x$, TKN, and Cl (Figures S10a, S10c, and S10f). In addition, WRTDS-K appears to be less biased than WRTDS in most cases for TP, SRP, and especially SS flux (Figures S10b, S10d, and S10e).

4.5. Model Application to Chesapeake Tributaries

In a pilot application, we implemented WRTDS-K (with optimal $\rho$) on water-quality data in the rivers draining to Chesapeake Bay. The goal of this effort was to evaluate the difference of the estimated fluxes from WRTDS-K and WRTDS, which is already used for flux estimation at these sites. These differences are evaluated at daily (Figures S11–S14), monthly (Figures S15–S18), and annual (Figures S19–S22) time scales.

We illustrate the differences using a case study of NO$_x$ in the Potomac River (Figure 9). On the daily scale, concentration estimates from WRTDS appear to follow the pattern of regression toward the mean and fail to capture the extreme values (Figure 9a). Notably, the WRTDS estimates were consistently lower than the observations for several months after the peak concentration observed in early 1996, which is associated with one of the biggest flood events during the period of water-quality data collection. Concentration estimates of the regular WRTDS model were persistently below the true values for a couple of months after the flood event. This suggests that the flood event (caused by a large regional rainfall and warm weather that happened at a time of a large regional snowpack) may have mobilized nitrate sources that are not typical of winter conditions. This anomaly, of several months’ duration, is captured well by WRTDS-K but not by the regular WRTDS model.
The differences between the WRTDS and WRTDS-K estimates depend on the temporal scales at which estimates are made (Figures 9b–9d). On the daily scale, the two sets of estimates have considerable differences, but generally scatter around the 1:1 line (Figure 9b). Differences in daily estimates produced a MAD (equation (11)) of 5,470 kg/day for the entire period. At a monthly scale, the two methods have fewer large differences (Figure 9c), and thus a smaller MAD (4,613 kg/day) than daily estimates. At an annual
scale, the two sets of estimates are almost always on the 1:1 line (Figure 9d), corresponding to a MAD (2,738 kg/day) that is even smaller than that of the monthly estimates. This scale dependence was also observed with all the other site-constituent combinations (Table S4)—that is, in all cases, the MAD value gets smaller from the daily scale to the monthly scale (i.e., $\text{MAD}_d/\text{MAD}_m > 1$) and gets smaller from the monthly scale to the annual scale (i.e., $\text{MAD}_m/\text{MAD}_y > 1$).

Overall, these results show that WRTDS-K and WRTDS estimates tend to be more comparable at coarse scales (e.g., annual) and that they become more distinguishable on short-term scales (e.g., daily), which is related to the residual adjustment in WRTDS-K. Such short-term based differences may have important implications to the modeling, research, and management efforts that rely on riverine flux estimates, including the calibration of Chesapeake Bay watershed and estuarine models, and exploration of terrestrial drivers of riverine flux patterns.

### Table 2
Percent Improvement in the Model MNSE Value for Daily Concentration Estimates by WRTDS-K (With Optimal $\rho$) Relative to WRTDS

| Constituent | S1 | S2 | S3 | S4 | S5 |
|-------------|----|----|----|----|----|
| NO$_x$      | 56 | 60 | 69 | 72 | 119|
| TP          | 13 | 20 | 17 | 19 | 36 |
| TKN         | 23 | 34 | 29 | 37 | 61 |
| SRP         | 16 | 20 | 21 | 28 | 45 |
| SS          | 10 | 14 | 11 | 11 | 20 |
| Cl          | 21 | 30 | 31 | 31 | 59 |

Note. For simplicity, the median improvement of the nine Lake Erie/Ohio River sites is presented.

Figure 9. Comparison between WRTDS and WRTDS-K (with optimal $\rho$) for NO$_x$ estimation in Potomac River: (a) daily concentration time series in a selected period (years 1995 and 1996), (b) comparison of daily flux estimates, (c) comparison of monthly flux estimates, and (d) comparison of annual flux estimates. The black line in (b)-(d) is 1:1 line. The points in (a) indicate days with observed concentrations.
5. Conclusions and Recommendations

This work provided statistical evidence to support the hypothesis that WRTDS daily estimates of concentration and flux can be improved by accounting for the autocorrelation structure of model residuals. Our results show that the WRTDS-K method generally produces more accurate estimates of daily concentration and flux values than WRTDS for all constituents, especially NO₃. The degree of improvement by WRTDS-K is strongly affected by the underlying sampling scenario; WRTDS-K estimates tend to be more accurate with increased water-quality sampling because more residuals can be exploited in the AR1 interpolation procedure. The performance of WRTDS-K depends on the selection of an appropriate $\rho$ value, which can vary depending upon the constituent being estimated and the sampling scenario being deployed. Model estimates were used to recommend optimal $\rho$ values for different water-quality constituents and sampling scenarios. In a pilot application, the optimal $\rho$ values were used to quantify nutrient and sediment fluxes from major tributaries to Chesapeake Bay, where it was shown that estimates from the two approaches were comparable at relatively coarse scales (e.g., annual) but became more distinguishable on short-term scales (e.g., daily).

Overall, the WRTDS-K method has the potential for broad applications for estimating concentration and flux from sparsely collected water-quality monitoring records, particularly those from geological and hydrological settings similar to the Lake Erie and Ohio River region. From a monitoring perspective, our results also demonstrate that flux estimation can benefit from both increased sampling frequency and targeted stormflow sampling. Software to compute WRTDS-K daily estimates of concentration and flux are available at http://usgs-r.github.io/EGRET/articles/WRTDS-K.html. This software is designed as an add-on to the existing EGRET software and it also includes some graphics capabilities to help visualize the differences between the original WRTDS estimates and the WRTDS-K estimates.

While this work demonstrates the potential of WRTDS-K, we recognize several areas that are appropriate for future research. First, it would greatly expand the applicability of the approach if the AR1 coefficient can be directly estimated from water-quality data, which are typically irregularly sampled. Second, an alternative approach is to add a cross-validation procedure in WRTDS-K to identify the optimal $\rho$ for any specific data set and use that $\rho$ value to run WRTDS-K. However, this may be difficult to implement because of the low sampling frequencies of traditional water-quality monitoring data. Third, we assumed in this work that WRTDS model residuals follow an AR1 model, but that may not always be the most appropriate model. In this regard, future efforts may consider more complex time series models such as the autoregressive moving average model or the autoregressive fractionally integrated moving average model. However, scientific advancement is needed to directly estimate the model coefficients for these models for the irregularly sampled water-quality data and to develop the analytical approach to fill the gaps between adjacent samples. Finally, as high-frequency monitoring sites and records become more available (Fovet et al., 2018; Halliday et al., 2015; Outram et al., 2014; Pellerin et al., 2012; Pellerin et al., 2016), WRTDS-K has the potential to estimate missing days when the sensors become inoperable, but this would require further research to evaluate the usefulness of WRTDS-K and its optimal settings for varying gap lengths within continuous records. Moving forward, scientists and researchers need to advance the understanding on these aspects and to translate the understanding to software tools that can be broadly applied.

Acknowledgments
This work was supported by the USEPA under grant “EPA/CBP Technical Support 2017” (07-5-230480) and by the USGS Water Mission Area. We would like to thank the late Tim Cohn (USGS) for providing the R code to implement the interpolation procedure under AR1. National Center for Water Quality Research is acknowledged for maintaining the high-quality data sets for the Lake Erie/Ohio River sites. Water-quality data for the other sites and river discharge data for all sites used in this work are available from the USGS National Water Information System database (U.S. Geological Survey, 2018). R scripts developed in this work are available from the corresponding author. The authors declare no conflict of interest. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. This is contribution number 5711 of the University of Maryland Center for Environmental Science.

References
Appling, A. P., Leon, M. C., & McDowell, W. H. (2015). Reducing bias and quantifying uncertainty in watershed flux estimates: The R package loadflex. Ecosphere, 6(12). https://doi.org/10.1890/ES14–00517.1
Aulenbach, B. T., & Hoover, R. P. (2006). The composite method: An improved method for stream–water solute load estimation. Hydrological Processes, 20(14), 3029–3047. https://doi.org/10.1002/hyp.6147
Bird, D. L., Groffman, P. M., Salice, C. J., & Moore, J. (2018). Steady-state land cover but non-steady-state major ion chemistry in urban streams. Environmental Science & Technology. https://doi.org/10.1021/acs.est.8b03587
Bowes, M. J., Smith, J. T., Jarvie, H. P., & Neal, C. (2008). Modelling of phosphorus inputs to rivers from diffuse and point sources. Science of the Total Environment, 392(2–3), 125–138. https://doi.org/10.1016/j.scitotenv.2008.01.054
Chanat, J. G., Moyer, D. L., Blomquist, J. D., Hyer, K. E., & Langland, M. J. (2016). Application of a weighted regression model for reporting nutrient and sediment concentrations, fluxes, and trends in concentration and flux for the Chesapeake Bay Nontidal Water-Quality.
Murphy, J. C., Hirsch, R. M., & Sprague, L. A. (2014). Antecedent flow conditions and nitrate concentrations in the Mississippi River basin. *Hydrology and Earth System Sciences*, 18(3), 967–979. https://doi.org/10.5194/hess-18-967-2014

National Center for Water Quality Research (2015). Tributary data download. https://www.heidelberg.edu/tributary-data-download.

Outram, F. N., Lloyd, C. E. M., Jonczyk, J., Benskin, C. M. W. H., Grant, F., Perks, M. T., et al. (2014). High-frequency monitoring of nitrogen and phosphorus response in three rural catchments to the end of the 2011–2012 drought in England. *Hydrology and Earth System Sciences*, 18(9), 3429–3448. https://doi.org/10.5194/hess-18-3429-2014

Park, Y. S., & Engel, B. A. (2015). Analysis for regression model behavior by sampling strategy for annual pollutant load estimation. *Journal of Environmental Quality*, 44(6), 1843–1851. https://doi.org/10.2134/jeq2015.03.0137

Pellerin, B. A., Bergamaschi, B. A., & Horsh bourgeois, J. S. (2012). In situ optical water quality sensor networks - workshop summary report. *Open-File Report* 2012–1044, 13 pp. U.S. Geological Survey, Reston, VA.

Pellerin, B. A., Stauffer, B. A., Young, D. A., Sullivan, D. J., Bricker, S. B., Walbridge, M. R., et al. (2016). Emerging tools for continuous nutrient monitoring networks: Sensors advancing science and water resources protection. *Journal of the American Water Resources Association*, 52(4), 993–1008. https://doi.org/10.1111/1752-1688.12386

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0; URL http://www.R-project.org/.

Stackpoole, S. M., Stets, E. G., Clow, D. W., Burns, D. A., Aiken, G. R., Aulenbach, B. T., et al. (2017). Spatial and temporal patterns of dissolved organic matter quantity and quality in the Mississippi River basin, 1997-2013. *Hydrological Processes*, 31(4), 902–915. https://doi.org/10.1002/hyp.11072

Stackpoole, S. M., Stets, E. G., Clow, D. W., Burns, D. A., Aiken, G. R., Aulenbach, B. T., et al. (2017). Spatial and temporal patterns of dissolved organic matter quantity and quality in the Mississippi River basin, 1997-2013. *Hydrological Processes*, 31(4), 902–915. https://doi.org/10.1002/hyp.11072

Stets, E. G., Lee, C. J., Lytle, D. A., & Schock, M. R. (2018). Increasing chloride in rivers of the conterminous U.S. and linkages to potential corrosivity and lead action level exceedances in drinking water. *Science of the Total Environment*, 613–614, 1498–1509. https://doi.org/10.1016/j.scitotenv.2017.07.119

Strickling, H. L., & Obenour, D. R. (2018). Leveraging spatial and temporal variability to probabilistically characterize nutrient sources and export rates in a developing watershed. *Water Resources Research*, 54, 5143–5162. https://doi.org/10.1029/2017WR022220

U.S. Environmental Protection Agency (2010). Chesapeake Bay Total Maximum Daily Load for Nitrogen, Phosphorus and Sediment, December 2010. Annapolis, MD: U.S. Environmental Protection Agency.

U.S. Geological Survey (2018). USGS water data for the Nation: U.S. Geological Survey National Water Information System database. https://doi.org/10.5066/F7P5KJN. Last accessed December 5, 2018.

Van Meter, K. J., Basu, N. B., & Van Cappel len, P. (2017). Two centuries of nitrogen dynamics: Legacy sources and sinks in the Mississippi and Susquehanna river basins. *Global Biogeochemical Cycles*, 31, 2–23. https://doi.org/10.1002/2016GB005498

Vedchia, A. V. (2003). Relation between climate variability and stream water quality in the continental United States. *Hydrological Science and Technology*, 19(1), 77–98.

Verma, S., Markus, M., & Cooke, R. A. (2012). Development of error correction techniques for nitrate-N load estimation methods. *Journal of Hydrology*, 432–433, 12–25. https://doi.org/10.1016/j.jhydrol.2012.02.011

Warner, S., Kiley, G., Morgan, G., & O’Halloran, J. (2009). Does quantifying antecedent flow conditions improve stream phosphorus export estimation? *Journal of Hydrology*, 378(1–2), 97–104. https://doi.org/10.1016/j.jhydrol.2009.09.009

Zhang, Q., & Ball, W. P. (2017). Improving riviereine constituent concentration and flux estimation by accounting for antecedent discharge conditions. *Journal of Hydrology*, 547, 387–402. https://doi.org/10.1016/j.jhydrol.2016.12.052

Zhang, Q., & Blomquist, J. D. (2018). Watershed export of fine sediment, organic carbon, and chlorophyll-a to Chesapeake Bay: Spatial and temporal patterns in 1984–2016. *Science of the Total Environment*, 619-620, 1066–1078. https://doi.org/10.1016/j.scitotenv.2017.10.279

Zhang, Q., Brady, D. C., Boynton, W. R., & Ball, W. P. (2015). Long-term trends of nutrients and sediment from the nontidal Chesapeake watershed: An assessment of progress by river and season. *Journal of the American Water Resources Association*, 51(6), 1534–1555. https://doi.org/10.1111/1752-1688.12327

Zhang, Q., Harman, C. J., & Kirchner, J. W. (2018). Evaluation of statistical methods for quantifying fractal scaling in water-quality time series with irregular sampling. *Hydrology and Earth System Sciences*, 22(2), 1175–1192. https://doi.org/10.5194/hess-22-1175-2018

Zhang, Q., Hirsch, R. M., & Ball, W. P. (2016). Long-term changes in sediment and nutrient delivery from Conowingo dam to Chesapeake Bay: Effects of reservoir sedimentation. *Environmental Science & Technology*, 50(4), 1877–1886. https://doi.org/10.1021/acs.est.5b04073