Environmental Modelling & Software 73 (2015) 148–166

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

A bootstrap method for estimating uncertainty of water quality trends

Robert M. Hirsch [a, *], Stacey A. Archfield [a], Laura A. De Cicco [b]

[a] U.S. Geological Survey, 432 National Center, USGS, Reston, VA 20192, USA
[b] U.S. Geological Survey, 8505 Research Way, Middleton, WI 53562, USA

ARTICLE INFO

Article history:
Received 23 July 2015
Accepted 30 July 2015
Available online 27 August 2015

Keywords:
Water quality
Bootstrap
Trend
Uncertainty analysis

ABSTRACT

Estimation of the direction and magnitude of trends in surface water quality remains a problem of great scientific and practical interest. The Weighted Regressions on Time, Discharge, and Season (WRTDS) method was recently introduced as an exploratory data analysis tool to provide flexible and robust estimates of water quality trends. This paper enhances the WRTDS method through the introduction of the WRTDS Bootstrap Test (WBT), an extension of WRTDS that quantifies the uncertainty in WRTDS-estimates of water quality trends and offers various ways to visualize and communicate these uncertainties. Monte Carlo experiments are applied to estimate the Type I error probabilities for this method. WBT is compared to other water-quality trend-testing methods appropriate for data sets of one to three decades in length with sampling frequencies of 6–24 observations per year. The software to conduct the test is in the EGRETci R-package.

Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Software

The statistical procedures presented here are all based on the Weighted Regressions on Time, Discharge, and Season (WRTDS) approach to water quality data analysis. The WRTDS is implemented in the EGRET (Exploration and Graphics for RivEr Trends), R-package (open source) available from the Comprehensive R Archive Network http://cran.r-project.org/web/packages/. The new software that implements the WRTDS Bootstrap Test (WBT) described in this paper is also an R-package called EGRETci, also available from the Comprehensive R Archive Network.

1. Introduction

More than 40 years after the passage of the Clean Water Act in the United States, large public investments and significant regulatory actions continue to be made in order to continue making progress towards the goals set forth in the Act (Knopman and Smith, 1993; Copeland, 2006). Public officials, land owners, and the general public express concern over perceived deterioration of water quality and seek to determine the magnitude of the impact that public and private investments and regulatory actions are having on the attainment of water quality goals (Broussard et al., 2012; National Research Council, 2011; Mehan, 2012) in order to decide about investing in further actions. On-going evaluations of the direction and magnitude of water quality trends remains an important task to support the achievement of water quality goals.

Various statistical methods have been used for more than 30 years to explore and analyze temporal trends in water quality. More recently, these methods have advanced as a result of several factors: increased lengths of consistent data sets, improvements in statistical methods, improvements in computer software and hardware, observations of a wide range of multidecadal trends in water quality, and improved understanding of watershed-based and in-channel processes affecting water quality. Examples of some of these methods include: Richards and Baker (2002), Langland et al. (2007), Ryberg et al. (2014), and Corsi et al. (2015). A part of these advancements has been the introduction of new approaches that stem from exploratory data analysis and smoothing concepts, including adaptation of locally weighted scatterplot smoothing (LOESS) (Cleveland and Devlin, 1988), and generalized additive models (GAMs) (Wood, 2006) to surface water quality data (see for example Reckhow and Qian, 1994; Langan et al., 2001; Morton and Henderson, 2008; and Hirsch et al., 2010). These methods are primarily aimed at a desire to characterize the timing, magnitude, and general nature of the trends observed.

Not surprisingly, there is an interest among many water quality professionals to have descriptions of trends be accompanied by statements of statistical significance, including confidence intervals

* Corresponding author.
E-mail addresses: rhirsch@usgs.gov (R.M. Hirsch), sarch@usgs.gov (S.A. Archfield), ldecicco@usgs.gov (L.A. De Cicco).

http://dx.doi.org/10.1016/j.envsoft.2015.07.017
1364-8152/Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

on the amount of change observed (e.g. Boesch et al., 2005). This interest is very legitimate. For example, the analysis may say that the mean concentration of nitrogen at a given monitoring site has increased by 1 mg/L over the past 30 years. Recognizing that typical monitoring strategies may only sample 6 to 12 times per year, one can expect that the estimate of 1 mg/L change is highly uncertain. If the analyst can state that the 90 percent confidence interval around that value ranges from 0.9 to 1.1 mg/L this relatively narrow range of uncertainty should provide a much stronger basis for action as compared to a result which states that the 90 percent confidence interval runs from −0.5 mg/L to 2.5 mg/L. This latter result suggests that although the likely direction of change is positive, there is actually a non-trivial chance that concentrations have not increased over the 30-year period. In this case, decision-makers may be inclined to exercise more caution in committing public or private resources to remedy the situation.

Whereas the need for such confidence interval estimates and associated statements of attained significance levels is great, it is not a simple matter to provide such estimates when the method of analysis is an exploratory approach that makes very few assumptions about the statistical properties of the data. This paper delivers an approach to adding uncertainty analysis to one particular exploratory data analysis method: Weighted Regressions on Time, Discharge, and Season (WRTDS) (Hirsch et al., 2010). We use a bootstrap (Diaconis and Efron, 1983; Efron and Tibshirani, 1994) procedure to provide complimentary uncertainty information along with the graphical and numerical outputs already provided by the WRTDS method. We call this the WRTDS Bootstrap Test (WBT). This paper briefly reviews the WRTDS method, and then describes the WBT. The bootstrap procedure used here is a new type of block bootstrap designed to account for the influence of serial correlation on the test results without attempting to explicitly model the correlation structure. Modeling the serial correlation of these kinds of water quality data sets can be very problematic given the relatively sparse and often irregular sampling that is common to such data sets. The block bootstrap approach introduced here approximately preserves the serial correlation for lags on the order of weeks to months and thus achieves Type I error rates that are relatively close to the nominal Type I error rate.

The block bootstrap approach is evaluated using a set of Monte Carlo simulations to estimate the Type I error probability (probability of detecting a trend when a trend was not present) as compared to the nominal significance level for this method under the null hypothesis that water quality conditions have not changed over the period of analysis. Type II error (the probability that a trend is present but not detected), although of great importance, was not evaluated here because of the multitude of different possible manifestations of departures from the null hypothesis that are possible. These include different rates of change, step functions versus ramp functions, and trends driven by point source changes versus those driven by non-point sources. WRTDS is designed to be sensitive to a variety of different types of trend scenarios, whereas most of the more common types of trend tests assume a simple and rather rigid model of the trend. Thus it is reasonable to assume that the WRTDS method will have an advantage in terms of Type II errors for a wide range of trend scenarios, but some disadvantage when the trend scenario postulated adheres closely to the assumptions around which other tests were designed. The wide range of trend scenarios would add greatly to the complexity of this study and may not be very illuminating. Thus, we kept our inquiry to the narrower question: is the WBT test accurate in terms of Type I error? The Monte Carlo simulations are based on three different generating models for discharge and concentration that are designed to replicate the statistical properties seen in actual water quality records. The Type I error probability resulting from the WBT is compared to the Type I error probability resulting from three common alternative trend analysis procedures: these are a multiple regression approach, the Seasonal Kendall test on residuals from a flow–concentration relationship, and the Seasonal Kendall test adjusted for serial correlation. These Monte Carlo simulations are further used to provide a suggested block-length for the test. Lastly, an example data set is evaluated using the WBT and several approaches for communicating uncertainty are presented. Because the WRTDS method is fundamentally an exploratory data analysis method, software that is relatively fast and interactive is crucial to the effective use of the method. Addition of the WBT analysis to define uncertainties has the potential to slow down that rapid interactive process. The desire to obtain the uncertainty information in a timely manner motivates the particular pathway this software development follows: aimed at providing useful uncertainty information without greatly slowing the overall analytical process. Hence the WBT uses some novel approaches to maximize computational speed. An important aim of this paper is to demonstrate (through Monte Carlo testing) that these approaches do not significantly compromise the validity of the test.

2. Overview of WRTDS method

The motivations for the WRTDS method and details of its computational techniques are described in Hirsch et al. (2010) and Hirsch and De Cicco (2014); many implementation details are omitted here in the interest of brevity. New notation and explanations of the method not published previously are presented throughout Section 2 in order to provide the concepts and mathematical symbology needed to explain the uncertainty analysis presented in Section 3. The WRTDS method has been implemented within an R package, know as EGRET (Exploration and Graphics for RivEr Trends) and is available on the Comprehensive R Archive Network (CRAN) at http://cran.r-project.org/web/packages/.

Major features of WRTDS include the following:

- It can detect and describe temporal trends that may not conform to linear or quadratic functional forms.
- It is suitable for use with irregularly spaced data.
- It does not assume that the discharge versus concentration relationship has the same shape throughout the period of record.
- It does not assume that the concentration residuals are homoscedastic.
- It does not assume that the seasonal pattern remains the same over the period of record.
- It can assess both concentrations and fluxes, recognizing that the trends in each of these measures of water quality can be quite different and even of different sign.
- It can not only provide estimates of the time series of annual mean concentrations and fluxes, but also time series of “flow-normalized” mean concentrations and fluxes which integrate over the probability distribution of discharge to remove the effect of interannual streamflow variability.

2.1. WRTDS estimation of daily concentration

The WRTDS model utilizes the sampled water quality data from an individual sampling site, along with the daily mean discharge at that site for the sampling dates, to develop an estimate of the concurrent daily mean concentration given by:

$$E[c_{ij}] = w(Q, T|Q_{ij}, T_{ij}) \tag{1}$$

where

$E[c_{ij}]$ is the expected value of concentration (in mg/L), on day i of year j
$w(Q,T)$ is a smooth continuous function of two variables, discharge (Q) in m^3/s, and time (T) in years,
$w(Q, T|Q_{ij}, T_{ij})$ is the function $w(Q,T)$ evaluated at Q_{ij} the observed daily mean discharge value for day i of year j, and T_{ij} the time value associated with day i of year j

The function $w(Q,T)$ can never be known exactly, but it can be estimated using the weighted regression approach of WRTDS. This estimate is denoted as $\widehat{w}(Q, T)$. Note that Q is the daily mean discharge on the day the sample was collected and not the instantaneous discharge at the time of sampling. One can use that function and the actual history of Q values over the period of record to estimate concentrations for every day.

$$\widehat{c}_{ij} = \widehat{w}(Q_{ij}, T_{ij}) \tag{2}$$

Where $\widehat{c}_{ij}$ is the WRTDS estimate of concentration for day i of year j. It is designed to be an unbiased estimate and experience has shown it to be very nearly unbiased over many applications. The period of record is defined here as the period over which there are a set of water quality samples (typically with no time gaps greater than two years and at least 200 observations) and for which there are daily mean discharge values for every day.

This function $\widehat{w}$ can be visualized as a contour plot such as the example shown in Fig. 1.

This function is evaluated at a set of regularly spaced grid points on a surface defined by time and log(Q), extending just beyond the observed range of Q and T values that were observed in the period of record. A large portion of that surface is depicted in Fig. 1. Individual estimates of concentration for a specific day and year, and the discharge that occurred that day, denoted $\widehat{c}_{ij}$, are determined by bilinear interpolation from this surface.

The estimation of this surface is accomplished through the use of weighted least-squares regressions of the form:

$$\ln(c_{ij}) = \beta_o + \beta_1 \ln(Q_{ij}) + \beta_2 T_{ij} + \beta_3 \sin(2\pi T_{ij}) + \beta_4 \cos(2\pi T_{ij}) + \varepsilon_{ij} \tag{3}$$

where:

$ln(c)$ = natural log of concentration in mg/L
$ln(Q)$ = natural log of daily mean discharge in m^3/s
T = Time in years

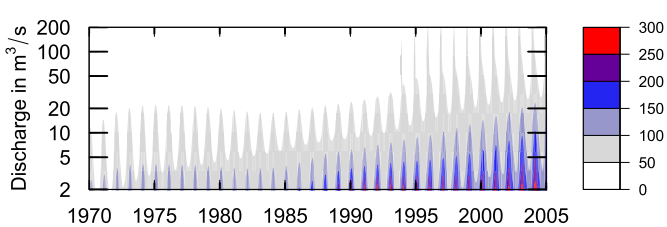

Fig. 1. WRTDS estimates of chloride concentration as a function of discharge and time, for the Milwaukee River at Milwaukee, WI. Concentrations are expressed in mg/L.

β_0, β_1, β_2, β_3, *and* β_4, are the regression coefficients
ε_{ij} is the error term, which is assumed to be normal with mean equal to zero and variance, σ^2, that varies smoothly across all of the values of $ln(Q)$ and T.

There is a separate weighted regression model for each grid point (Q,T) where the weight on each observation in the data set is the product of three separate weights related to the "distance" of the sample point from the grid point in dimensions of time, log(Q), and season of the year. Note that the "grid points" are a regular array of combinations of Q and T values, equally spaced in the ln(Q) and T dimensions. The "sample points" are defined in terms of the same set of coordinates, but are located at the particular values of ln(Q) and T at which the sample was taken. The estimates at the grid points use equation (3) and a weighted subset of the sample data points where the subset and the weights are determined by the "proximity" of the sample point to the selected grid point. The method is implemented in a manner that accommodates censored data by using weighted Tobit regression (Tobin, 1958) as an alternative to weighted least squares regression regardless of the presence of censoring in the given data set. Note that the regression coefficients (the 4 $\widehat{\beta}$ values and $\widehat{\sigma}$) are smooth functions of Q and T, but for notational brevity this functional dependence is not explicitly indicated.

The estimate of c for any given value of Q and T is:

$$\widehat{c} = \widehat{w}(Q, T) = \exp\left\{\widehat{\beta_o} + \widehat{\beta_1}\ln(Q) + \widehat{\beta_2}T + \widehat{\beta_3}\sin(2\pi T) + \widehat{\beta_4}\cos(2\pi T) + \frac{\widehat{\sigma^2}}{2}\right\} \tag{4}$$

The last term in equation (4) is needed to prevent the problem of re-transformation bias.

2.2. WRTDS estimation of flow-normalized concentration

The portion of the variation in concentration that is due to streamflow variation creates a great deal of noise in the annual time series of average annual estimated concentrations, making it very difficult to accurately assess changes that represent true progress (or lack of progress) in improving water quality. The particular sequence of discharges during the period of record can also introduce spurious trends, which can come about due to a persistent period of either high or low flow at either the beginning or end of a water quality record. WRTDS uses an approach called flow normalization (FN) that filters out the influence of the inter-annual variations in streamflow to produce a time series of "flow normalized concentrations" (FNC). FNC is a representation of concentration that integrates over the probability distribution of discharge in order to remove the effect of year-to-year variation in discharge.

The FNC for day i of year j (denoted here as c^*_{ij}) is defined as:

$$c^*_{ij} = \int_0^\infty w(Q, T_{ij}) \cdot g_{ij}(Q) dQ \tag{5}$$

where $g_{ij}(Q)$ is the probability density function (pdf) of Q (discharge) specific to day i of year j.

Neither of the two functions in this equation can ever be known exactly, so both must be estimated. The estimate of w (called $\widehat{w}$) is estimated by the WRTDS weighted regression method described

above. For the estimate of the pdf of discharge, $g(Q)$, the WRTDS flow-normalization process depends on an assumption that discharge for any given day of the calendar year is stationary over the period of record. Thus, it should not be applied where management actions that substantially modify streamflow have taken place over the period of the water quality record. These actions could include building or removing a large dam, major changes in consumptive water use, major changes in baseflow due to groundwater depletion, or major changes in land drainage. Human-induced climate change is another potential source of nonstationarity in streamflow, but at present the magnitude and direction of this change is highly uncertain. Future versions of the WRTDS method will consider streamflow as a nonstationarity random variable, but for now, the FN method should be restricted to situations where any such changes are relatively minor compared to the other drivers of water-quality change and sources of uncertainty.

The estimate used for the pdf of discharge is denoted as $\widehat{g}_i(Q)$. It is not indexed by year (j) because of the assumption of stationarity of discharge. But it is indexed by day (i), because it is specific to the day of the year. Rather than using a complex statistical model to represent the distribution of discharge for a given day, a nonparametric approach is used. The estimate $\widehat{g}_i(Q)$, for day i is defined by the observed values of Q for that day of the calendar year, and the probability of each of those values is equal to $1/n_y$ where n_y is the number of years of observed discharge values for that day of the year. The entire daily discharge record (for example in a 20-year record this would be about 7300 values) is used to estimate the daily frequency distributions.

Using the estimates $\widehat{w}$ and $\widehat{g}(Q)$, we can compute the set of estimates, $\widehat{c}^*_{ij}$, of the flow-normalized concentration for each day of the period of record.

$$\widehat{c}^*_{ij} = \int_0^\infty \widehat{w}(Q, T_{ij}) \cdot \widehat{g}_i(Q) dQ \tag{6}$$

Equation (6) could be expressed as a sum rather than an integral because the $\widehat{g}_i(Q)$ function is represented as a set of point masses rather than a continuous function. However we express it as an integral to emphasize its relationship to equation (5), which is properly expressed as an integral. Before proceeding further, a comment should be made about the notation with respect to day and year. The simplifying assumption being made here in the description of the method is that all years are 365 days long and the time period being analyzed is the calendar year. The actual computational method used in the EGRET R code accommodates leap years, see (Hirsch and De Cicco, 2014, Appendix 2, page 18). Also in EGRET, the time period for which averages and totals are computed are not required to be calendar years. They can be water years (the default in EGRET), or seasons made up of contiguous months, or even individual months. For simplicity of presentation in this paper the time period considered is calendar years, and in the Monte Carlo simulations presented later in the paper they are water years.

2.3. Trend characterization in WRTDS

The period being evaluated for trend covers a set of years starting with y_s, the starting year, and ending with y_e, the ending year. The full period of record can extend prior to y_s and extend beyond y_e or it can be just limited to the time span from y_s to y_e.

The true mean value of FNC for y_s is:

$$FNC_s = \frac{1}{365} \cdot \sum_{i=1}^{365} c^*_{is} \tag{7}$$

and the true mean value of FNC for y_e is:

$$FNC_e = \frac{1}{365} \cdot \sum_{i=1}^{365} c^*_{ie} \tag{8}$$

where c^*_{is} and c^*_{ie} are the true values of FNC on day i of years y_s and y_e respectively.

Using the approach described above, we can make estimates of FNC for each day in the period of record ($\widehat{c}^*_{ij}$), and these estimates can be substituted for true concentrations in (7) and (8) to obtain estimated mean FNC values $\widehat{FNC}_s$ and $\widehat{FNC}_e$.

The true change in FNC over the trend period is Δ^*_c:

$$\Delta^*_c = FNC_e - FNC_s \tag{9}$$

Placing this in the context of classical hypothesis testing, the null hypothesis (H_{c0}) is that there is no change in FNC between year s and year e. Formally, we could say:

$$H_{c0} : \Delta^*_c = 0 \tag{10}$$

The alternative hypothesis, H_{c1}, is that they are not equal, $H_{c1} : \Delta^*_c \neq 0$. Note that the alternative is two-sided, FNC can either have increased or decreased over the trend period.

Using the WRTDS method the estimate of the change over the trend period is defined as:

$$\widehat{\Delta}^*_c = \widehat{FNC}_e - \widehat{FNC}_s \tag{11}$$

The obvious, but not unique, circumstance under which H_{c0} could be true would be the case where both the w function for y_s and y_e are identical and the g functions for any given day of the year in year y_s is identical to the g function for that same day of the year in y_e for all 365 days of the year. For the purposes of the Monte Carlo simulations of the null hypothesis (described in section 5.1) these two functions are simulated as stationary throughout the simulated period, not just equal at the start year and end year.

The WRTDS method not only considers trends in concentration, but also trends in flux. The two variables (concentration and flux) are tightly related, and the computations for both in WRTDS are tightly linked. However, the changes that take place in one variable may be quite different from changes in the other. Annual average concentration is a time average of daily concentrations and hence the concentration on the days of very high discharge have the same influence on the average as do days of moderate or low discharge. In contrast, annual average flux is dominated by conditions that happen on the days of the highest discharge and the concentrations on days of very low discharge are relatively inconsequential to the annual average flux. It is entirely possible that average concentrations could decline over a period of years, because of reductions in point source contributions of a pollutant, but average fluxes could rise because of increases in non-point source inputs that happen primarily on high flow days. Given the potential for different results for flux trends as compared to concentration trends, the WRTDS model explicitly models flux but does so in a manner that is consistent with its approach to concentration.

The flow-normalized flux (FNF) in kg/d for day i of year j (denoted here as f^*_{ij}) is defined as:

$$f_{ij}^* = 86.4 \cdot \int_0^\infty Q \cdot w(Q, T_{ij}) \cdot g_{ij}(Q) dQ \tag{12}$$

which is the same as the definition for FNC for that day and year, except that the integrand is multiplied by discharge and by a unit conversion factor (86.4).

The true mean value of FNF for y_s is:

$$FNF_s = \frac{1}{365} \cdot \sum_{i=1}^{365} f_{is}^* \tag{13}$$

and the true mean value of FNF for y_e is:

$$FNF_e = \frac{1}{365} \cdot \sum_{i=1}^{365} f_{ie}^* \tag{14}$$

where f_{is}^* and f_{ie}^* are the true values of FNF on day i of years y_s and y_e respectively.

In a similar manner to FNC, define the true change in FNF over the trend period is Δ_f^*:

$$\Delta_f^* = FNF_e - FNF_s \tag{15}$$

And the null hypothesis (H_{f0}) that there is no change in FNF between y_s and y_e is:

$$H_{f0} : \Delta_f^* = 0 \tag{16}$$

The alternative hypothesis, H_{f1}, is that they are not equal. $H_{f1} : \Delta_f^* \neq 0$.

Using the WRTDS method, the estimate of the change over the trend period is $\widehat{\Delta}_f^*$, which is defined as:

$$\widehat{\Delta}_f^* = \widehat{FNF}_e - \widehat{FNF}_s \tag{17}$$

If the w functions for y_s and y_e are identical and the g functions for y_s and y_e are identical, then it follows that both H_{c0} and H_{f0} are true. However, because the strength of statistical evidence for trend in FNC can be quite different from the strength of the statistical evidence for trend in FNF it is appropriate to conduct hypothesis tests on both. There are a number of possible outcomes that can arise from such testing: (a) rejecting both null hypotheses, with both of them indicating increase, both indicating decrease, or one indicating an increase and the other a decrease; (b) rejecting one of the null hypotheses and not the other; or (c) rejecting neither.

3. Estimation of uncertainty and hypothesis testing in WRTDS

Here we consider how we might evaluate the two hypotheses and estimate the uncertainty in the estimates of change in FNC and FNF over the trend period.

3.1. Application of the bootstrap to WRTDS

When statistical methods are complex, such as the smoothing procedure applied in WRTDS, it is generally not feasible to make statements about the uncertainty of results using simple mathematical expressions such as those that apply to ordinary regression. Bootstrapping is a common approach to the problem of describing the uncertainty of these more complex analyses. This approach was first introduced by Efron (1979) and is discussed in general terms by Diaconis and Efron (1983) and in detail in by Efron and Tibshirani (1994). The latter text (pages 70–80) as well as (Efron, 2005) specifically describe the application of the bootstrap to weighted least-squares regression smoothing algorithms such as loess (locally weighted scatterplot smoothing). Similar principles can be used in applying it to WRTDS, which is a specific multidimensional application, similar in many respects to loess. Bootstrap techniques have been used previously in studies of hydrologic processes, see for example: Rajagopalan and Lall (1999) and Ames (2006), including studies of constituent transport in rivers, see for example: Aulenbach and Hooper (2006), Rustomji and Wilkinson (2008), Ide et al. (2012), and Vigiak and Bende-Michl (2013). It has also been applied to water quality trend analysis by Darken et al. (2000, 2002).

Given the complexity of the processes that give rise to the time series of concentrations observed in a river, we assume, just as with all bootstrap applications, that the observed data provides a reasonable representation of the population behavior for the period over which the null hypothesis will be evaluated. In bootstrapping, we re-use our data many times over to represent the kind of variability that we can expect in the real population. In the simplest form of bootstrapping, if we have N observations of concentration over our period of record, we would use a random sampling algorithm to select bootstrap replicates from the data set. A bootstrap replicate has the same number of observations as the actual data set (N observations) and they are selected from the sample data set, with replacement. That means that in any one of the bootstrap replicates we may find that the actual observation from a given day might, or might not, be present in that replicate, or it may be present multiple times. To be used with the WRTDS model we must select the observations in our bootstrap replicates to preserve their particular location both in time (year and season) and in discharge. In other words, every sample selected is a vector of values (c, T, *and* Q).

The bootstrap method used here is a type of block bootstrap which has the effect of approximately maintaining the short term serial correlation structure that exists in the data set but without having to attempt to model that serial correlation structure. The serial correlation structure of concern here is that of the residuals (ε_{ij} in equation (3)). Residuals from WRTDS models when plotted as a function of time, tend to have runs of positive (or negative) values that persist for several days to several weeks. See Hirsch and De Cicco, 2014, (fig. 34) for an example of this tendency for residuals to have long sequences of positive (or negative) residuals. In some cases these runs may reflect "event-related" characteristics wherein a particular high discharge event has high concentrations, more than would be expected from the overall model estimates, and another high discharge event has lower than expected concentrations. Estimating the correlation structure of the residuals is a difficult problem because most water quality data sets are irregularly spaced in time and when viewed as a daily time series the majority of the days are missing values. In those instances where there are daily time series that are either complete or have only a small number of missing values, we typically find that there is a modest amount of serial correlation in the residuals out to lags of a month or two. Darken et al. (2002) provides some examples of the serial correlations observed at one or two month lags for eight analytes at about ten sampling sites in the eastern US (after removal of seasonal and discharge related sources of variation). The overall mean of the lag 1-month correlation coefficients in their study was approximately 0.10 and at lag 2-months it was 0.07. Examples of lag-1 day correlations were investigated by Lettenmaier (1976) who showed correlations for a number of sites and variables on the order of 0.85 and demonstrated that the auto-correlation function shows more persistence than would be expected from an AR(1) model at a daily time step. A common approach to accounting for serial correlation, but without the need to estimate a specific time

series model is to use a block bootstrap approach, see (Efron and Tibshirani, 1994; and Politis and Romano, 1994). Therefore, rather than treating each observation in the record as if it were an independent event, the samples are treated in groups that may be many weeks in duration.

Our block bootstrap method uses a sampling block that is based on an interval of time and not based on a number of samples. It is the time-domain analogy to the spatial "Grid-based block bootstrap" method proposed by Lahiri and Zhu (2006) which was designed for irregularly spaced samples. This approach is used because many water quality sample records have periods of dense and sparse sampling. The approach prevents the bootstrapping procedure from oversampling of the denser sampled periods of the record but tends to keep intact the set of samples from individual high or low discharge events (because their residuals are likely to show substantial serial correlation). This time-based, rather than sample-based block bootstrap was developed for the WBT out of concerns about highly unequal sampling frequencies (and a focus on intensive event sampling) that are common to these types of records. This method preserves the dependent nature of these closely spaced samples.

The time-based block bootstrap resampling algorithm used here works as follows:

1) A block length (B), such as 100 days, is selected by the analyst. The choice of block length is discussed in section 4.1.3.
2) A time frame of days that serve as potential starting dates for individual blocks is established. This time frame starts with the day that is $B-1$ days before the first sample in the data set and runs to and includes the last sampled day in the data set. For example, if $B=100$, and the first sample were 1979-10-05 and the last sample were 1987-09-27 then the days in this time frame would be the days 1979-06-28 through 1987-09-27.
3) Randomly select (with replacement) a day (call it day G) from this time frame.
4) Establish a time window that runs from day G through day $G+B-1$ (inclusively) and select all of the sample values in that time window for use in constructing the bootstrap replicate. Note that some of these windows may include no sample values. Also note that some of the days within the time window may fall before the start of the record or after the end, and thus for those parts of the record no samples are selected (because they do not exist). In the case described in step 2, if G were 1979-10-03, then there would only be 98 possible sample dates considered in selecting water quality samples and the group of samples taken would be only those samples from the first 98 days of the sampled period.
5) Repeat steps 3 and 4, adding samples to the bootstrap replicate. Call the total number of samples in the replicate N_r. When $N_r \geq N$ (where N is the number of samples in the true data set) stop taking bootstrap replicates and trim the set of replicates by deleting enough from the end of the last selected group so that $N_r = N$.

3.2. Hypothesis testing using the block bootstrap in WRTDS

Each time a bootstrap replicate (of N observations) is selected it can then be subjected to the same WRTDS analysis that the observed data set was subjected to, and a set of bootstrap replicate trend results can be computed and saved. Of course, because of the bootstrapping process the data that are used to estimate the WRTDS model for the replicate may have some sample values repeated two or more times (the same concentration on the same date) while other sample values are not used to estimate the model. To improve computational efficiency, the WBT only requires that the $w(Q,T)$ function be estimated for the two years, y_s and y_e, so that the difference between the two years can be computed. However, all of the concentration and daily discharge data from the entire bootstrap replicate are used in the estimation. For purposes of this discussion, we will consider only the estimate of the trend in the FNF between a particular pair of years (e.g. the difference between the FNF estimate from 1992 and from 2012), but the exact same steps are undertaken with FNC using the same bootstrap replicates that are used for the calculation of FNF. The estimate of trend in FNF between any two years is a standard output of the existing WRTDS procedure. The WBT determines, using the bootstrap method, a 90% confidence interval on the magnitude of that trend in FNF, and decides if H_{f0}, the null hypothesis that the trend in the expected value of FNF is zero, should be rejected. Our test is designed to have a probability of Type I error of 0.1, hence the use of the 90% confidence interval. The confidence interval computations use the approach described by Davison and Hinkley (1997) known as "basic bootstrap confidence limits." This particular type of bootstrap confidence interval was selected because of its computational simplicity and its accuracy in the face of highly asymmetric data distributions.

For the *kth* replicate the difference in FNF between the y_e and y_s is denoted D_{fk}. It is computed from the bootstrap sample in the same manner as $\widehat{\Delta}_f^*$ is computed from the actual sample (using equation (17)). The kth bootstrap estimate of change in flux is Δ_{fk} where

$$\Delta_{fk} = 2 \cdot \widehat{\Delta}_f^* - D_{fk} \tag{18}$$

This equation is derived from Davison and Hinkley (1997, equation (5.6), p. 194).

Using the kth bootstrap replicate, we also make similar computations for concentration resulting in the kth bootstrap estimate of change in concentration, denoted Δ_{ck}. The bootstrap sampling is repeated M times and for each bootstrap replicate the values of Δ_{ck} and Δ_{fk} are computed.

3.3. An adaptive Bayesian approach to determine the number of bootstrap replicates (M)

In many bootstrap applications, the analyst determines in advance the number of bootstrap replicates that will be selected. Because WRTDS is computationally intensive it is helpful to the analyst to bring the bootstrap replicates process to a rapid conclusion when it is clear that having more replicates will be unlikely to change our decision to reject or not reject the null hypothesis. To accomplish that goal of minimizing the number of replicates, the following adaptive Bayesian approach is used to determine the number of bootstrap replicates (M).

This adaptive bootstrap procedure is designed to produce two classical hypothesis testing results: 1) reject or, fail to reject, H_{c0} and 2) reject or, fail to reject, H_{f0}. Each test provides a two-sided p-value associated with the null hypothesis and confidence intervals for Δ_c^* and for Δ_f^*. The algorithm uses an α level of 0.1 for each of the tests, corresponding to a Type I error probability of 0.1. The value of α equal to 0.1, rather than the more common choice of 0.05, was selected for two reasons. The first is that we are concerned about the power of the test, meaning that we want to have a high probability of identifying real trends when they exist and we are willing to trade off a somewhat higher risk (larger α value) that we might declare that there is a trend (reject H_0) even when no trend exists. The other is a practical consideration for this bootstrap procedure. Setting the Type I error rate (α) to a lower value, such as 0.05 or 0.01, would greatly increase the computational burden for the test.

The adaptive algorithm described here takes some prudent shortcuts that allow us to stop running additional bootstrap replicates when they don't materially add value to the result. This includes two kinds of conditions. One is the case where the statistical support for rejecting both of the null hypotheses is very weak. The other is the case where the statistical support for rejecting both of the null hypotheses is very strong.

We use an adaptive Bayesian approach to iteratively provide a metric of the support for the null hypotheses after each new bootstrap replicate is created and evaluated. We will start the discussion of this approach exploring the hypothesis trend for FNF. In the test, this analysis is simultaneously carried out for FNC. Then, using results from the accumulated bootstrap replicate results for FNC and FNF, a set of criteria are used to determine if another bootstrap replicate should be generated or if there is sufficient information to stop the procedure. There is extensive literature on stopping rules. See for example, Sanborn and Hills (2014) and the many references therein. The stopping rule used here is designed for the case where two hypotheses are being tested and where their test statistics are correlated with each other (in this case the test for trends in FNC and FNF). The Monte Carlo experiments described in Section 4 provide a demonstration that the adaptive stopping rule results in a Type I error rate for the WBT that is approximately correct.

Let us define π_f as the fraction of bootstrap replicates in an infinite number of bootstrap replicates for which the estimated change in FNF from y_s to y_e is positive ($\Delta_{fk} > 0$). Of course, under the null hypothesis, H_{f0}, $E[\pi_f] = 0.5$. That is, the expected number of replicates showing an increase in FNF is half of the total number of replicates.

At any stage in the bootstrap process, we can make an estimate of π_f, which we can denote as $\widehat{\pi}_f$. It is defined as the mean of the Bayesian posterior distribution of π_f using a non-informative prior. We use the commonly-implemented Jeffreys' prior (Jeffreys, 1998), which is a Beta distribution having both shape parameters equal to 0.5. The Beta distribution is a commonly used prior distribution in Bayesian analysis for binomial proportions (Congdon, 2007), which is precisely the application in use here. This value of 0.5 for the shape parameters is the standard value used for the non-informative prior (see Gelman et al., 2014, p. 53). Of interest is the proportion of the replicates that are positive, indicative of an upward trend. Using this prior, it follows that the posterior mean is $\widehat{\pi}_f = (x_f + 0.5)/(M + 1)$ where x_f is the number of positive changes (the number of bootstrap replicates for which $\Delta_{fk} > 0$) and M is the number of bootstrap replicates. Based on the resulting posterior distribution determined from M replicates, we can estimate a credible interval for the true value, π_f. We will denote the lower and upper bounds of this Bayesian credible interval as π_{fL} and π_{fU} respectively such that $Prob(\pi_{fL} < \pi_f < \pi_{fU}) = 1 - \alpha_p$. Thus, α_p is the probability of a Type I error for the true value of π_f, falling outside the specified credible interval. It should not be confused with α which is the probability of a Type I error regarding the trend in FNF. The credible interval used here is a central interval, which means that the tail areas of the posterior distribution are equal. The posterior probability distribution of π_f given the observed values of x_f and M is a Beta function with the first shape parameter equal to $x_f + 0.5$ and the second shape parameter equal to $M - x_f + 0.5$. The central interval is based on this posterior Beta function defined so that $Prob(\pi_f < \pi_{fL}) = \alpha_p/2$ and $Prob(\pi_f > \pi_{fU}) = \alpha_p/2$. The estimates of π_{fL} and π_{fU} are provided by the binom.bayes function in the binom R-package. Based on a set of Monte Carlo experiments α_p was set equal to 0.3. Using this value means that there is a probability of 0.7 that the true value of π_f lies within the interval and a 0.15 chance that it lies above and a 0.15 chance that it lies below the interval. This α_p value was chosen based on the goals of achieving the approximately correct overall α level ($\alpha = 0.1$) for each of the two tests (one for FNC and one for FNF) and also limiting the average number of replicates required to complete the test under the null hypothesis. Monte Carlo experiments were used during the development of the WBT to aid in selecting an appropriate value of α_p. Using smaller values of α_p added to computational time but had a negligible effect on accuracy of the overall test.

3.3.1. First stopping criteria

This Bayesian approach is used to guide the stopping criteria for the bootstrap procedure. One possible outcome of this adaptive bootstrap procedure is that after some number of replicates we find that the entire credible interval lies between 0.05 and 0.95. That is, the bootstrap results (which are that there are x_f positive values in M replicates) indicates that $0.05 \leq \pi_f \leq 0.95$ with probability $(1 - \alpha_p)$ or greater. For example, we may have carried out only 10 replicates and we find that among the 10 estimates of change (Δ_{fk}) there are 6 with a positive value and 4 with a negative value. Given this information, we can state with a very high level of certainty that π_f is not less than 0.05 and not greater than 0.95 and we can stop the bootstrap replicate process and decide that we should not reject H_{f0}. For this example, if $\alpha_p = 0.3$ the credible interval for π_f would be (0.44, 0.74) and this would be sufficient basis to conclude that we should not reject the null hypothesis, H_{f0}. Even if we chose to be very risk adverse about this decision and set $\alpha_p = 0.01$ the credible interval for π would be (0.23, 0.90) suggesting that it is highly unlikely that $\pi_f > 0.95$ or that $\pi_f < 0.05$. The algorithm enables the process to end when the evidence suggests that π_f is not close to 0 or 1.

The procedure, as implemented, also imposes a minimum number (M_{min}) of replicates requirement. The procedure will only stop if $M \geq M_{min}$. The lowest value allowed for M_{min} in the WBT software is 9. This value of M_{min} is used in the simulations described in this paper unless otherwise noted. The user is free to set $M_{min} > 9$ and generally doing so will increase the precision of some of the other results of the process beyond the simple decision to reject or not reject the null hypothesis, but at a cost of increased computer time. The suggestion that $M_{min} > 9$ is based on simulations that indicated that the selection of a lower minimum creates an unnecessarily large probability of Type II error. We return to discussions of the influence of M_{min} in Sections 5.2 and 5.4.

3.3.2. Second stopping criteria

In addition to having a stopping criterion for cases where π_f clearly lies between 0.05 and 0.95, there is a stopping criterion for cases where π_f clearly lies well below 0.05 or well above 0.95. If $\pi_{fL} > 0.95$ or $\pi_{fU} < 0.05$ the bootstrap process should stop. This procedure also imposes a minimum number of replicates requirement, $M \geq 31$. An example of a situation where this rule gets applied could be the following: if $\alpha_p = 0.3$ and there are 0 positive values of Δ_{fk} in 31 replicates. In this situation the credible interval for π_f would be (0, 0.017) and this would be sufficient basis to stop the process because it is highly unlikely that additional replicates would cause us to conclude that $\pi_f > 0.05$. Another example would be 1 positive value in 53 replicates, the credible interval would be (0.007, 0.049) and this would also be sufficient basis to stop the process. The minimum value of 31 was selected based on the recognition that when H_0 is rejected it is likely the case that the analyst will be asked to provide an approximate p-value for the test result. The p-values computed (as discussed in Section 3.4) will be highly imprecise with sample sizes smaller than 31.

3.3.3. Simultaneous testing of FNF and FNC

The two stopping criteria for FNF (described above) are also applied to FNC after each bootstrap replicate is run. The algorithm

specifies that the process should stop only if one of the two stopping criteria is met for both FNF and FNC. If one of the stopping criteria is met for only one (FNF or FNC) or for neither, then the bootstrap replicates process continues and both the FNC and FNF criteria are reevaluated until both of them meet one of the stopping criteria. Computationally there is very little burden to computing both FNC if FNF is being computed because the bulk of the computation is related to estimating $w(Q,T)$ which is required for estimates of both. So, both statistics are computed in subsequent iterations even though the stopping criterion for that statistic has been met. Experience using the test indicates that the estimates of π_f and π_c at any given stage of the bootstrap process are often quite different from each other (meaning that the strength of statistical support for stationarity of flux can be quite different from that for concentration). This observation suggests that α_p does not need to be as low as 0.1 or 0.05 in order to assure that the overall test procedure has a Type I error probability close to 0.1. The results of the Monte Carlo experiment (discussed below) demonstrate that this combination of parameters used in the stopping criteria result in a reasonably accurate Type I error rate.

3.3.4. Maximum number of replicates to be drawn

Finally, there is an additional rule that when M reaches M_{max} the bootstrap process will stop. The default value in the code is $M_{max} = 100$ but users are free to increase or decrease it. Increasing it will improve the precision of the results, again at a cost in terms of computer time. If $M = M_{max}$ the algorithm will stop when that replicate is completed, regardless of the other criteria described above. The selection of all of the parameters involved in the stopping rule is a trade-off between speed of computation and reliability of the results. If the analyst is running many analyses and has the time and computer resources, then setting $M_{min} = M_{max} = 100$ will provide results that are much more reproducible and precise than if M_{min} *and/ or* M_{max} were set to lower values. But, if the test is viewed as an extension of an interactive EDA process, then smaller values of these parameters may be warranted. As an example, in a data set of 776 observations, and a trend period for which the data do not provide strong evidence of a trend in either concentration or flux, the choice of using $M_{min} = 9$ versus $M_{min} = 100$ (along with $M_{max} = 100$) resulted in computational times of about 1.2 min and 9.2 min respectively (on a Macintosh OS X version 10.9.5, with a 2.5 GHz Intel Core i5).

3.4. Determination of confidence intervals in WBT

The conclusions about the existence of a trend and the sign of the trend are determined on the basis of the 90% confidence interval. This process is used regardless of the criteria that led to stopping the bootstrap replications. The 90% confidence interval on the trend in FNF is computed by interpolating the 5% and 95% quantiles of the sample cumulative distribution function of the bootstrap estimates, Δ_{fk}, using the Weibull plotting position (Stedinger et al., 1993). The lower bound of this confidence interval is Δ_{fL} and the upper bound is Δ_{fU}. The conclusion that there is a trend is based on the following rule: if $(\Delta_{fL} \cdot \Delta_{fU}) > 0$, then we conclude that there is a trend (reject H_{f0}). In other words, if the lower and upper confidence limits are of the same sign, we conclude that there is a trend. If they are of opposite signs (meaning that they straddle the value of 0), then we regard the test as inconclusive. The direction of the trend is based on the sign of the WRTDS FNF trend estimate ($\widehat{\Delta}_f^*$). The analogous computation is made for concentration: including estimates of the upper and lower bounds on the 90% confidence interval (Δ_{cL}, Δ_{cU}), conclusions about the existence of a trend (rejecting H_{c0}), and conclusions about the direction of the trend, based on the sign of the WRTDS FNC trend estimate ($\widehat{\Delta}_c^*$). Note that it is possible for the sign of the trend in FNC can be opposite that for FNF.

In null-hypothesis significance testing it is common to report the p-value alongside the decision to reject or not reject H_0. The p-value gives an immediate sense of the strength of the evidence in support of this decision and provides a measure of the analyst's confidence in the result. The WBT can produce a statistic that is the functional equivalent to the p-value. For the set of bootstrap replicates (Δ_{fk}, $k = 1, 2, \ldots M$), we use the sample cumulative distribution function of the Δ_{fk} values evaluated at $\Delta_f = 0$ to determine p_{0f}. Linear interpolation is used to compute a p_{of} value at $\Delta_f = 0$. The two-sided p-value for FNF is then computed as $P_f = 2 \cdot min(p_{of}, 1 - p_{of})$. In those cases where all $\Delta_{fk} > 0$ or all $\Delta_{fk} < 0$, then we state the result as $P_f < 2/(m + 1)$. This provides a way to express the results of the WBT in terms of a p-value in addition to statements of the 90% confidence interval and the decision to reject or fail to reject H_{f0}. The two-sided p-value for FNC is computed in the same manner, based on the bootstrap replicate values Δ_{ck}, $k = 1, 2, \ldots M$.

4. Using Monte Carlo experiments to evaluate the WRTDS Bootstrap Test (WBT)

The discussion of the WBT to this point has not considered the effect of the block-length (B). In this section we develop a Monte Carlo experiment used to guide the selection of B. In addition, the results of this Monte Carlo experiment are used to evaluate how well the attained significance level of the test corresponds to the nominal significance level ($\propto = 0.1$). This Monte Carlo experiment also considers three other common trend analysis techniques and determines how well their attained significance level corresponds to their nominal significance level. Ultimately, the decision to use any given trend test procedure depends on two major considerations: One is the extent to which the test outputs can describe the nature and magnitude of the trend that may have occurred. The other is the degree to which it can accurately represent the level of uncertainty in the results, in the face of some of the complexities that exist in the data sets being analyzed. The Monte Carlo experiment applied here is designed to both provide guidance on the selection of B, and also provide a basis for assessing the accuracy of the Type I error rate for the WBT in comparison to some other common tests. At the end of this section we briefly consider the accuracy of the Type I error in the face of varying amounts of censoring in the data set.

4.1. Design of the Monte Carlo experiment

4.1.1. Water quality and streamflow data sets

The experiment is based on three stationary stochastic simulation models with a range of characteristics that are representative of some of the important properties of water quality data in large watersheds. The models are designed using three actual data sets that have daily or close to daily sampling for a period of a decade or more. This high frequency of sampling is needed to provide a basis for reasonable definition of the serial correlation properties of the concentration data in conjunction with the seasonality and covariation of concentration with discharge.

The three models are based on the following actual data sets:

1. The chloride record for the Cuyahoga River at Independence, OH (designated here as CUYA). At the data collection site, this is a 1836 km^2 watershed. It is estimated to have been 34% urban in 1992, and 40% urban in 2006 (Corsi et al., 2015). Chloride is an important pollutant in this watershed because of the extensive use of sodium chloride to melt snow and ice on pavement in this

significantly urbanized watershed. The water quality data were collected by the National Center for Water Quality Research at Heidelberg University in Tiffin, OH and downloaded from their web site http://www.heidelberg.edu/academiclife/distinctive/ncwqr/data/data. The discharge data are from the USGS, station 04208000. The data used are from the version of this database as updated on 2014-10-08. The portion of the record used to develop the model covers water years 2004–2013 and the stationary model used in the Monte Carlo simulations is based on conditions for calendar year 2008. The water quality record used consists of 4087 samples collected over this 3653-day period (2004–2013). The data set was thinned to contain no more than one sample per day (by deleting all but the last sample of the day). The resulting chloride data set consists of 3359 samples, which means that there are 294 days with no data (92% of the days had at least one sample). The chloride concentrations in these 3359 samples had a minimum of 27.6 mg/L, a maximum of 1170 mg/L, and a median of 146 mg/L. The USEPA chronic water quality criterion for chloride is 230 mg/L.

2. The second is the dissolved nitrate record from the Vermilion River, at Pontiac, IL, (designated here as VERM) a highly agricultural watershed of 1500 km^2. The water quality data set was collected at a frequency of one sample per day for over 12 years by the Northern Illinois Water Company and provided to us by Professor Momcilo Markus, University of Illinois and Illinois State Water Survey. The discharge data used are from USGS station 05554500. The watershed and data set are described by Sogbedji and McIsaac (2006). For purposes of developing the model the period of 2769 consecutive days from 1991-10-01 to 1999-04-30 were used. The nitrate record is representative of nitrate issues faced in many watersheds in the corn-belt region of the US due to the contributions of reactive nitrogen to the rivers and reservoirs of the region and ultimately to the Gulf of Mexico. The nitrate concentrations in the record range from 0.1 mg/L to 26.0 mg/L, with a median value of 8.8 mg/L. These concentrations are somewhat typical of the corn-belt region of the US and are much higher than levels observed in non-agricultural areas or areas with other cropping patterns (Dubrovsky et al., 2010).
3. The third data set is total phosphorus for the Maumee River at Waterville, OH (designated here as MAUM). The watershed at this site is 16,400 km^2 and the land use is 90% agriculture, 1% urban and the remainder largely woodland. The water quality data were collected by the National Center for Water Quality Research at Heidelberg University in Tiffin, OH and downloaded from their web site http://www.heidelberg.edu/academiclife/distinctive/ncwqr/data/data. The data used are from the version of this database as updated on 2014-10-08. The discharge data are from the USGS, station 04193500. The watershed is a major source of water and phosphorus to the western basin of Lake Erie and it has been well documented (Han et al., 2012; International Joint Commission, 2014) that the phosphorus derived from this river is crucial to the creation of cyanobacter blooms in Lake Erie which can result in the presence of algal toxins in the Lake. The data used to create the MAUM model consists of 4165 concentration values from samples collected during water years 2006 through 2013. For purposes of estimating the relationship of concentration to discharge and time, all 4165 concentrations were used. For estimating the serial correlation structure only one observation per day was used (the last observation of the day) and this resulted in a data set of 2780 days of observations out of a total of 2922 possible days. The median concentration was 0.183 mg/L and the maximum was 1.182 mg/L.

4.1.2. Stochastic streamflow and water quality models

These three data sets provide the basis for creating the three stochastic stationary models that are used in the Monte Carlo experiment. In all three cases it is likely that the actual data sets are non-stationary, but the process of building the model is designed to create a realistic but stationary model based on the behavior observed in the data. The stochastic model developed from each of these data sets consists of four parts: 1) a stationary model of the seasonality of discharge (representing the log of discharge as a function of sine and cosine of time of year), 2) a model of the serial correlation structure of the daily residuals from this stationary discharge model (represented by a low order ARMA model), 3) a stationary model of water quality concentrations as a function of discharge and time of year (represented in the same form as a WRTDS model), and 4) a model of the serial correlation structure of the daily residuals from this water quality model (using a low order ARMA model).

The stationary models were created to provide realistic coupled time series of discharge and water quality data. They are not intended to be the best possible statistical characterizations of the discharge and concentration data for each case. They are intended to create reasonable stationary data sets for testing the true Type I error rate of the WBT. In each case, the model is defined by the following: (The actual values of the various model coefficients are presented in appendix A and the properties of the models are represented graphically there as well).

A time series of simulated log daily discharges for each day ($i = 1, 2, \ldots, N_s$) is generated as follows:

$$ln(Q_i) = \beta_o + \beta_1 \cdot \sin(2\pi T_i) + \beta_2 \cdot \cos(2\pi T_i) + \sigma \cdot \varepsilon_i \tag{19}$$

Where

N_s is the number of days in the simulation period
$ln(Q_i)$ is the natural log of discharge (m^3/s) on day i in the simulation period
T_i is the decimal value of time, in years, associated with day i.
σ is the standard deviation of the log discharge around its seasonally varying mean value.
β_0, β_1, *and* β_2 are model coefficients, estimated by linear regression.
ε_i is generated from an ARMA(p,q) process, with mean = 0, standard deviation = 1, and normally distributed, with one value for each of the N_s days of the simulation period.

Based on examination of the actual data sets, the order of the ARMA process was established as the ARMA model with the lowest residual error variance, for which all of the parameters were significantly different from zero (based on the ratio of the parameter estimate to its standard error being greater than 2 in absolute value). One additional adjustment was made to the discharge values generated by this process. The $ln(Q_i)$ values are subjected to a linear transformation that is designed so that the minimum and maximum generated discharge values in the simulated record are equal to the minimum and maximum discharge values in the observed record that was used to fit the water quality simulation model (described below). This step is included so that the fitted water quality model can be used without having to extrapolate beyond the range of the observed data.

The water quality simulation model has the same structure as the WRTDS model, except that it is forced to be stationary. The way

it is created is to estimate the function $\mu(Q,T)$ which is the representation of the expected value of $ln(c)$ as a function of Q and T. Then, a particular calendar year near the center of the period of record is selected and a stationary (but seasonal) model $\mu^*(Q,T)$ is created by repeating that one year's segment of the function for all of the years in the simulation period. Thus, for any given Q and any given time of year (the fractional part of T) the expected value of $ln(c)$ is always the same. Similarly, the function $\lambda(Q,T)$ is estimated. It is a representation of the standard deviation of $ln(c)$ around the estimated mean $\mu(Q,T)$. Using the same selected calendar year, the function $\lambda^*(Q,T)$ is created by repeating $\lambda(Q,T)$ from that selected year for all of the years in the simulation period. The two functions $\mu(Q,T)$ and $\lambda(Q,T)$ are estimated from the full data set using the method described by Hirsch and De Cicco (2014).

A time series of $ln(c)$ values of length N_s days is generated as:

$$\ln(c_i) = \mu^*(Q_i, T_i) + \lambda^*(Q_i, T_i) \cdot e_i \tag{20}$$

where the e_i values are generated from an ARMA(p,q) process, with mean = 0, standard deviation = 1, and normally distributed, with one value for each of the N_s days of the simulation period. This ARMA(p,q) process is different from the one described above for discharge simulation. It is estimated from the time series of standardized residuals from the estimation of the functions $\mu(Q,T)$ and $\lambda(Q,T)$ from the actual data set. But, because some of the data sets used did not have a sample on every day, there are a small number of missing values in the time series of standardized residuals. One additional step is added to the simulation, which is a linear adjustment to the time series of e_i values so that their mean and standard deviation match those from the actual data set.

4.1.3. Approach to the selection of B, the block length

In the description of the WBT in section 3.2, the parameter B was introduced. It is the block length, expressed in days. If B were set to 1 day, then the procedure would be identical to a standard bootstrap approach in which individual samples are drawn, with replacement. If B were set to a very large value (say a thousand or more days in a data set with the length of 10 or 20 years) then it would greatly limit the variability of the bootstrap trend estimates. Note that in the limit, if B were set equal to the length of the total record of sampled days, the bootstrap replicates would be nearly identical to each other because the first randomly selected starting date would result in selecting a very large fraction of the whole data set and the next starting date would be very likely to be sufficient to obtain the necessary sample of size N. The ideal is to set B to a value that approximates the time-scale of the correlation structure observed in the water quality data. Because most long-term water quality records are very sparsely sampled (for example one sample every 30 or even 60 days) and are usually irregularly spaced in time, meaningful estimates of this structure is very difficult at best. The topic of the serial correlation structure of water quality data is an area that is beginning to see increased attention, particularly as the use of continuous water-quality sensors become more common (Kirchner, 2006; Kirchner and Neal, 2013). Sorting out the roles of seasonal variation, discharge-driven variation, and a wide range of possible forms of trend along side the role of serial correlation structure is a very challenging problem. The problem of selecting an appropriate block length for bootstrap methods on time series data is a difficult one, even with uniformly spaced data where serial correlation structure can be evaluated reasonably well (Efron and Tibshirani, 1994). Against this backdrop of issues, we used Monte Carlo simulations run with various values of B in order to determine the influence that the choice of B may have on the observed Type I error rate. We use these results to provide guidance on the selection of B by observing the extent to which the observed Type I error rate departs from the nominal Type I error rate.

The simulations used to evaluate the effect of block length (B) were done as follows. For each of the three models (CUYA, VERM, and MAUM) three record lengths were considered (10 years, 20 years, and 30 years). For the 20 and 30 year simulations three different sampling frequencies were simulated with uniform spacing: 6 per year, 12 per year, and 24 per year. For the 10-year record lengths only the frequencies 12 and 24 per year were considered. For a 10-year record, a 6 sample per year frequency would only result in 60 observations, which is a sample size for which WRTDS would not be an appropriate choice as a trend testing method. For each of these 24 combinations of models, record lengths and sampling frequencies 500 simulated records were produced. Each one of these records was tested for trend in FNC and FNF using 6 different block lengths (B values of 25, 50, 100, 200, 300, and 400 days). The observed overall Type I error rate for that individual case is denoted R_W, which is the sum of the cases where H_{c0} was rejected plus the number of cases where H_{f0} was rejected, divided by 1000 (the total number of tests conducted, 500 for FNC and 500 for FNF). The desired outcome is for $R_W \cong \alpha$ where $\alpha = 0.1$. Note that for this number of iterations, if we assume that $E[R_W] = 0.1$, the observed value of R_W would be in the interval (0.085–0.116) with a 90 percent probability assuming the tests for FNC and FNF are independent. If the two trend results were perfectly correlated this interval would be (0.078–0.122).

4.1.4. Use of three other trend tests in the Monte Carlo experiment

For purposes of evaluation of the attained significance level of the WBT, three other common tests used for water quality trends were also considered.

The first of these tests is referred to here as "ESTIMATOR" (denoted ESTIM). It is based on the ESTIMATOR model, which is a linear regression-based test that uses a quadratic representation of trend. This model has been used for many years to evaluate water quality trends at monitoring sites that are at the downstream ends of the major rivers draining to the Chesapeake Bay (Langland et al., 2007) and has been used in other studies such as an overview of water quality trends in the Missouri River Basin (Sprague et al., 2006). This test fits a regression model of the form:

$$\ln(c) = \beta_0 + \beta_1 \cdot \ln(Q) + \beta_2 \cdot (\ln(Q))^2 + \beta_3 T + \beta_4 T^2 + \beta_5 \cdot \sin(2\pi T) + \beta_6 \cdot \cos(2\pi T) + \varepsilon$$

The null hypothesis is that $\beta_3 = \beta_4 = 0$ and the test is based on the magnitudes and the degree of uncertainty about the two fitted coefficients β_3 and β_4 in relation to the overall unexplained variation. Details about the test are present in Langland et al. (2007, p. 17–19). Rejection of the null hypothesis is considered to be a conclusion that there is a "trend in flow-adjusted concentration" comparing y_s to y_e. Rejecting the null hypothesis indicates that for any given combination of time of year and discharge the conditional distribution of concentration has changed over time. In fact, the model is predicated on the assumption that the trend in $ln(c)$ is a quadratic function of time, and that function applies for any value of Q and for any time of year. The observed Type I error rate for this test is denoted R_E. This test is conducted 500 times for on the same data sets that were tested by WBT. If $E[R_E] = 0.1$, the observed value of R_E would be in the interval (0.078–0.122) with a 90 percent probability.

The second test is the Seasonal Kendall Test on residuals from the log concentration versus log discharge relationship (Hirsch et al., 1982). It is denoted here as SEAKEN and it was implemented using the rkt package in R. It applies the Seasonal Kendall test for trend using 12 "seasons", except in the case where there are

only 6 samples per year, in which case it is 6 "seasons" per year. In the case of 24 samples per year the year is divided into 12 seasons of equal length and the test is based on the median of the two values for each season of each year, rather than all 24 values for the year. In the implementation of the SEAKEN test used here, the Seasonal Kendall test is conducted on the residuals from a regression of the form:

$$\ln(c) = \beta_0 + \beta_1 \cdot \ln(Q) + \beta_2 \cdot (\ln(Q))^2 + \varepsilon.$$

The test is an intrablock method, (van Belle and Hughes, 1984) designed under an assumption that the form of the fitted discharge versus concentration model applies across seasons and across years and that the direction and magnitude of the trend is the same in all seasons.

Neither ESTIM nor SEAKEN take into account any serial correlation in the residuals from their respective models. As such, one can anticipate that for them the observed Type I error rate for this test (denoted R_S) and for ESTIM (R_E) will be greater than 0.10.

Finally the Seasonal Kendall Test on residuals, adjusted for serial correlation (SEAKENA) (Hirsch and Slack, 1984) is applied. The procedure is the same as for SEAKEN, but the computation of the variance of the Seasonal Kendall test statistic is modified based on the correlation in the ranks of the data across the various seasons. If all of the underlying assumptions are met (regarding the nature of the discharge versus concentration relationship and consistency of the trends across seasons) this test has been shown to have an observed Type I error rate (denoted R_{SA}) that is very close to α for data sets of 10-years or greater (Hirsch and Slack, 1984).

4.2. Results of simulation experiments

Consider first the MAUM model, with a record length of 20 years, and a sampling frequency of 24 samples per year (for a total of 480 samples). Fig. 2 depicts the results from the Monte Carlo simulation. What Fig. 2 indicates is that there is some difference in the observed Type I error rates for the tests of trend in FNC and FNF. Exploring these differences over the 24 simulations did not reveal any consistent relationship between their differences and the block length (B). The R_W values as a function of B are shown as the brown line in Fig. 2. The block length (B) which results in the R_W value closest to the desired level of 0.10 is at $B = 50$, but the differences across the B values from 50 through 400 are relatively minor. Any one of them results in a R_W value between 0.12 and 0.132. In contrast, R_E is much higher (0.154), R_S is similar to R_W, and finally R_{SA} is much closer to α, with a value of 0.09.

Overall, the results of this one case suggest the following: If we could, in fact, know that the record we wanted to evaluate for trend had properties similar to the MAUM model, then we could attain a reasonably accurate (in terms of Type I error) test using the WBT with B at any value in the range of 50–400 (and perhaps greater than 400). From the standpoint of accuracy of the Type I error rate, WBT it is a good deal better than ESTIM, slightly worse than SEAKEN and a good deal worse than SEAKENA. However, accuracy of Type I error rates should not be the only deciding factor in selecting a trend test. The WRTDS method, in conjunction with the WBT, offers several additional advantages over the other tests. These include the ability to distinguish between cases where trend in concentration may be weakly indicated while trend in flux is strongly indicated (or the opposite). This can be of critical importance based on the water quality objectives (instream quality versus delivery to a downstream water body) and in terms of distinguishing between probable

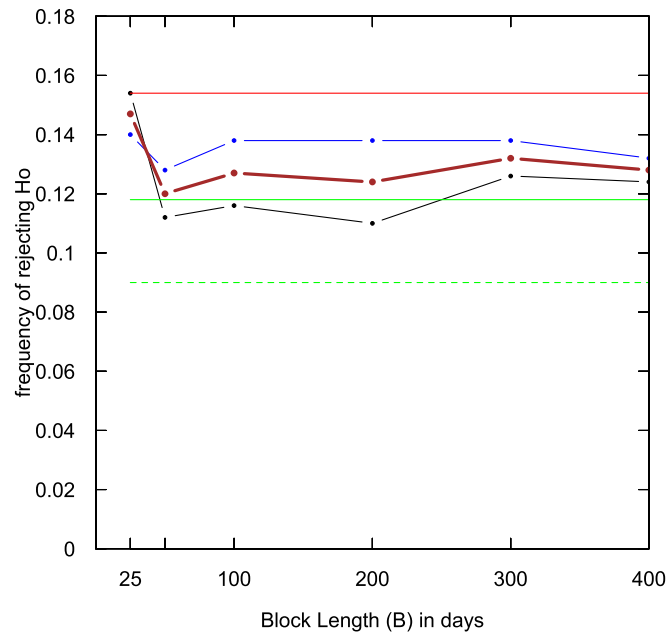

Fig. 2. Results of Monte Carlo simulation (500 iterations) using the MAUM model, with a record length of 20 years and a sampling frequency of 24 samples per year, showing frequency of Type I error as a function of block size (B). The black line shows the Type I error rate for FNC, the blue line is the result for FNF, the brown line is the average of the rates for FNC and FNF (R_W). The red line shows the Type I error rate for ESTIM, green is for SEAKEN, and dashed green is for SEAKENA. Note that the methods ESTIM, SEAKEN, and SEAKENA do not involve the use of block bootstrapping and thus the results are shown as horizontal lines on this figure, for purposes of comparison with the WBT (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

causes of the water quality changes (those dominant at low flow versus those dominant at high flow). WRTDS also provides a flexible characterization of the temporal pattern of the trend. In contrast, the ESTIM method can only describe the temporal pattern as a quadratic equation, which may or may not be truly representative of the temporal pattern of change. SEAKEN and SEAKENA make no attempt to define the temporal pattern, but simply provide a single valued indicator of trend slope, based on the Seasonal Sen Slope estimator (Hirsch et al., 1982). Thus, we are faced with a trade-off, with SEAKENA providing the most accurate Type I error rate but a very simplified and constrained description of the trend, versus WRTDS in conjunction with WBT providing a slightly biased Type I error rate but a much richer description of the nature of the trend, including the ability to distinguish between the trend in concentration and trend in flux. It is our conclusion that on balance, the better choice is WRTDS/WBT although confirmation of results using SEAKENA could be a useful check on the WRTDS/WBT results.

A second example is shown in Fig. 3. It is for the Monte Carlo simulation using the CUYA model with a record length of 30 years with 24 samples per year. In this case the value of R_W closest to α is at $B = 100$, and for that B value, R_W is 0.11. The value of R_E is 0.18 and R_S is 0.15, indicating rather large inaccuracies of the Type I error rates for the ESTIM and SEAKEN tests. The SEAKENA test performed very will with $R_{SA} = 0.10$. Overall the WBT performed quite well and that is true over a broad range of block lengths from 100 to at least 300 days.

4.2.1. Use of the Monte Carlo experiments to provide general guidance to the selection of B and evaluation of the WBT

The approach we used is this: For any of the 8 experimental designs (defined by record length and sampling frequency) we

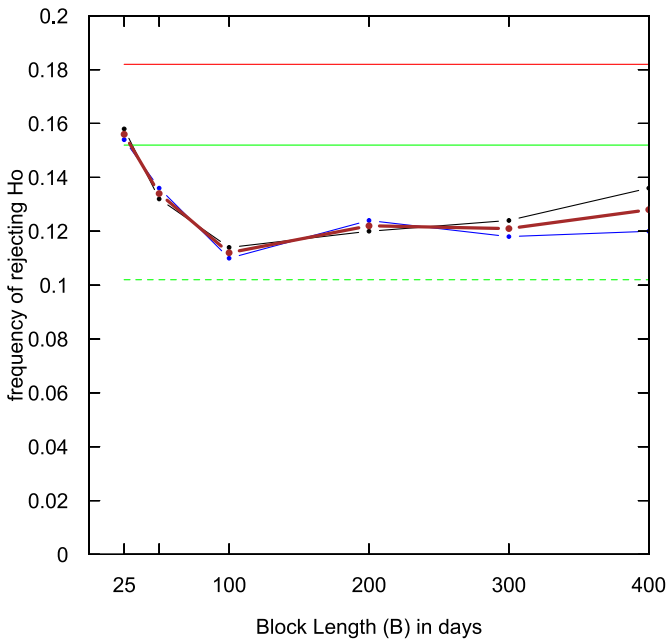

Fig. 3. Monte Carlo simulation (500 iterations) using the CUYA model, showing frequency of Type I error as a function of block size (B). The black line shows the Type I error rate for FNC, the blue line is the result for FNF, the brown line is the average of the rates for FNC and FNF (R_W). The red line shows the Type I error rate for ESTIM, green is for SEAKEN, and dashed green is for SEAKENA. Note that the methods ESTIM, SEAKEN, and SEAKENA do not involve the use of block bootstrapping and thus the results are shown as horizontal lines on this figure, for purposes of comparison with the WBT (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

produced a graphical representation of R_W for each value of B. Fig. 4 is a graphical representation of these results in this case for a record length of 20 years with 12 samples per year. Each symbol represents the value of one of the three simulations (CUYA, VERM or MAUM) for the given B value for a total of 18 points (3 models by 6

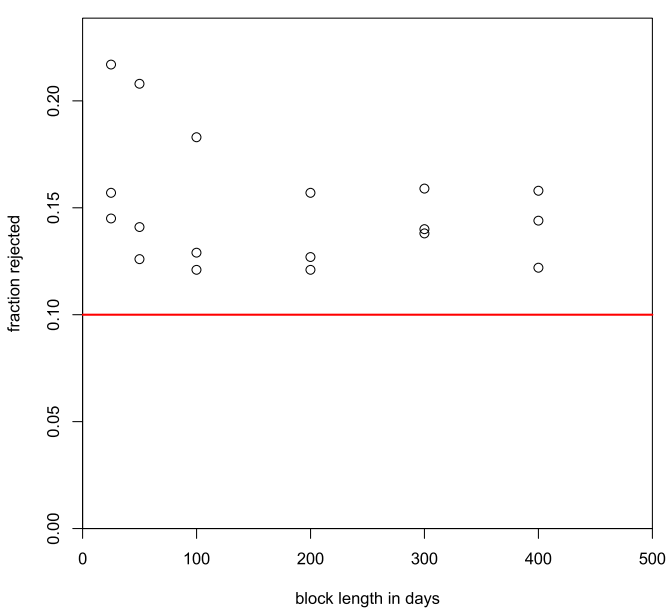

Fig. 4. Results of Monte Carlo simulation for record length 20 years, 12 samples per year. Each circle represents the observed Type I error rate (R_w) for the WRTDS Bootstrap Test for the three cases (CUYA, VERM, MAUM) and a specific bootstrap block length (B).

block lengths). A full set of these figures (8 in total, one for each experimental design) is provided in the supplemental material (Appendix B). Our working assumption is that the analyst will not have sufficient density of data to determine which of our three models best approximates their data set (in terms of serial correlation) so the guidance is based on finding the value of B that produces the most robust results across the set of models used.

These results were examined to see if there was a clearly optimal value of B for a given experimental design (record length and sampling frequency). We found no systematic differences in the shape of the relationship of B to R_W as a function of record length and/or sampling frequency. Thus, we lumped all of the simulation results and computed mean values of R_W (across all models, all record lengths, and all sampling frequencies) for each value of B. These values were 0.166, 0.149, 0.137, 0.132, 0.133, and 0.134 for B values of 25, 50, 100, 200, 300, and 400 respectively. Realistically, the mean values for block lengths of 200, 300, and 400 cannot be considered to differ from each other (they are based on 24,000 trials of the test, and with that sample size the width of the 95 percent confidence interval on the true value R_W is approximately 0.004). Given these results we have adopted 200 days as our choice for block length. But, no strong argument can be made against other values such as 300 or 400 days. Given the fact that the correlation time scale in our simulations (and presumably in actual concentration residuals) is about 100 days, then any block length greater than about 100 should be adequate.

4.2.2. Use of Monte Carlo experiment results to compare trend tests

Table 1 provides a summary of results for all of the tests applied for each experimental design (table entries are the mean value of the observed Type I error rates over all three models for each design). What Table 1 shows is that the WBT, using $B = 200$ generally has a Type I error rate that is slightly above the nominal rate of 0.10. This departure is somewhat larger for the longer record lengths. When compared against the other tests we see that the WBT error rate is closer to the nominal rate than is the ESTIMATOR test for the sampling frequencies of 12 and 24 per year and markedly so at 24 per year. Comparing with SEAKEN, WBT also has an error rate closer to the nominal rate for frequencies of 24 samples per year but for a frequency of 12 samples per year the results were mixed. At 30 years and 12 samples per year SEAKEN was notably closer to α than WBT, but at lower sampling frequencies the two tests had similar error rates. Finally, we see that SEAKENA consistently had the error rate closest to α and was, in all cases, not significantly different from 0.1.

As discussed above, if accuracy of Type I error rates were the only consideration in the selection of a test, then SEAKENA would be the preferred test overall. But, if the exploratory features of WRTDS are of interest to the user (representation of changing

Table 1
Results of Monte Carlo simulations. Mean values of observed Type I error rates across all three simulation models. Tests used are: WBT (R_W), ESTIM (R_E), SEAKEN (R_S), and SEAKENA (R_{SA}).

Record length	Frequency samples per year	Observed Type I error rates			
		R_w	R_E	R_S	R_{SA}
10	12	0.12	0.13	0.13	0.11
10	24	0.13	0.22	0.17	0.10
20	6	0.12	0.11	0.11	0.10
20	12	0.14	0.15	0.15	0.11
20	24	0.13	0.22	0.17	0.10
30	6	0.13	0.12	0.12	0.10
30	12	0.15	0.14	0.12	0.09
30	24	0.14	0.21	0.17	0.11
All eight cases		0.13	0.16	0.14	0.10

slopes of the trend over time, or differences between trends in concentration versus trends in flux) then WRTDS in conjunction with the WBT is a preferable approach.

4.3. Monte Carlo experiment to consider censored sampling

One additional Monte Carlo experiment was conducted to determine if the WBT is robust in the presence of censoring (i.e. where sample values are reported as less than some analytical limit). The Monte Carlo simulation was run using the MAUM case, with a sampling frequency of 12 samples per year for 20 years. The test was run 500 times using reporting limits that would result in an average of 5%, 10%, 25%, and 50% of the values as censored as well as a control run that had no censoring. The WBT estimated p-value for the no censoring case was 0.132. For the four censoring levels considered, the WBT estimated p-values were 0.135, 0.131, 0.130, and 0.134, respectively. The differences across censoring levels are not substantial. Therefore, one can conclude that the method is quite robust (in terms of Type I error probability) against even large amounts of censoring. This suggests that the WBT is appropriate for use with censored data, at least to the 50% censoring level and sample sizes of 240 or more observations.

5. Considerations regarding the presentation of uncertainty information from the WBT

There are a number of decisions for the analyst to make regarding how to conduct the WBT and how the outputs of the WBT results should be presented. These include: a likelihood-based approach to reporting results as an alternative to the null-hypothesis significance testing approach, consideration of the tradeoff between repeatability of results and the choice of numbers of replicates to run (which are a function of the M_{min}, *and* M_{max} values selected). The options for outputs include the use of graphical representation of the confidence region around the trend line, and the presentation of the distribution of bootstrap replicate outcomes to evaluate the probability of the true trend exceeding some selected magnitude (other than a magnitude of zero).

5.1. Communication of uncertainty results

The WBT described above follows the general approach known as Null-Hypothesis Significance Testing (NHST). NHST is considered by many to be a standard requirement for publication of almost any type of scientific result (such as the analysis of trend), but at the same time it has been subject to considerable criticism (see for example, Nicholls, 2001; Cohn and Lins, 2005; or Vogel et al., 2013) and adherence to the NHST approach varies across scientific disciplines. One concern relates to the arbitrary nature of the selected value of α, the significance level. Another concern is related to the highly constrained nature of the null hypothesis. Cohn and Lins (2005) characterized this by noting that rejection of the null hypothesis can arise because of a wide range of possible ways in which the natural system may depart from the null hypothesis, not just the existence of trend. This includes the potential for trend-like behavior to arise as a result of long-term persistence. Vogel et al. (2013) focus their criticism of NHST on the degree to which it is fixated on Type I error (the probability of rejecting the null hypothesis if it were actually true) versus concerns over Type II error. Type II error can be thought of as failure to recognize an important signal, when that signal should provide the basis for taking action.

Consider three examples. In one case a set of environmental actions have been put in place over a decade and we wish to know if, under those actions, progress is being made towards agreed-upon goals. The evidence provided by the data may not be convincing in the sense that it allows us to reject the null hypothesis in favor of the alternative that the trend in concentration (or flux) is downwards, and yet the evidence may be strong enough such that one could make a statement such as, "Even though we cannot reject the null hypothesis, we have high confidence that the data are indicating that water quality is improving and given the lag times involved in fully realizing the benefits of the actions taken, staying the course with the existing policies that have adopted are appropriate at this time and warrants a check back in a couple of years to re-evaluate if conditions are truly improving." The second example is one where the evidence shows that, under current policies, concentrations (or fluxes) have been rising in recent years even though the data do not provide sufficient evidence to allow us to reject the null hypothesis. A third case is one in which there is virtually no indication in the data that conditions are either improving or deteriorating and we wish to convey that there is no substantial support for any conclusion about the direction of the change.

What is needed is a lexicon that can be used to describe the degree of statistical support that the data set and the associated analysis provides, regarding the likelihood that the direction of change has been positive (or negative) over some specified period of time. We propose here a set of terminology partly based on language used by the Intergovernmental Panel on Climate Change (see Mastrandrea and Mach, 2011). The descriptive statements we propose are presented in Table 2. Adopting our terminology, one might say, for example: "it is highly likely that the trend in nitrate between 1990 and 2010 was upward" and this would mean that we believe that there is at least a 95 out of 100 chance that the direction was upward. Similarly in a case where we believe that there is at least a 66 out of 100 chance that the trend was upward we could say that it "is likely that the trend over the 1990 to 2010 period was upward." We argue here that these types of statements may be more suitable to the types of decisions that need to be made with respect to water quality. It provides decision makers with a formal, rather than subjective, evaluation of decisions to take an action or not, as they consider the tradeoffs involving the risks, costs, and benefits of various outcomes.

Making this type of statement is a very natural approach to the presentation of uncertainty information. For example, the decision makers may conclude that there is sufficient evidence that the current strategy is not working and that stronger action is warranted. Consider an example where WRTDS estimates a positive trend over some period of interest and the WBT results show an upward trend in 80 out of 100 cases. If the analyst used only a classical statistical NHST approach, they might report "we cannot reject the null hypothesis of no trend at $\alpha = 0.1$, and furthermore, the two sided p-value in this case is 0.4, suggesting very weak evidence of trend." However, using the likelihood descriptors proposed here, $\widehat{p} = 0.8$ which is in the range of 0.66–0.9 we could conclude that "it is likely that there is an increasing trend". The former classical type of statement would typically lead to

Table 2
Definitions for descriptive statements of likelihood of increasing trends for WRTDS Bootstrap Test (WBT) as a function of $\widehat{\pi}$, the posterior mean estimate of the probability of an increasing trend.

Range of $\widehat{\pi}$ values	Descriptors
$\geq$0.95 *and* $\leq$1.0	Highly Likely
$\geq$0.90 *and* <0.95	Very Likely
$\geq$0.66 *and* <0.90	Likely
>0.33 *and* <0.66	About as Likely as Not
>0.1 *and* $\leq$0.33	Unlikely
>0.05 *and* $\leq$0.1	Very Unlikely
$\geq$0 *and* $\leq$0.05	Highly Unlikely

complacency because it suggests that "we don't have strong proof of a growing problem" while the latter likelihood-style presentation says "we are pretty sure that conditions are not improving, and thus we need to step up our actions if we wish to improve water quality." The alternative approach proposed here is commonly associated with debate about the Bayesian versus frequentist views of the world (Press and Press, 1989).

The likelihood approach is more akin to what people actually use in their daily decision making process. They want to know if it is likely that something is true. Then, using that likelihood information along with an evaluation of the consequences of action or no action, they can make an informed decision about taking action. Seeking a very low p-value as a pre-requisite to action is not necessarily the most rational approach. However, it is a standard that has been used in science for many years. It may be appropriate for making claims that some "treatment" is actually efficacious in achieving the desired outcome (say in consideration of medical treatment or an agricultural practice designed to increase crop yields). These kinds of studies have the advantage that the experimenters can select a large sample size in order to give a high degree of certainty about the efficacy or lack of efficacy of a given treatment. But, in the context of evaluating the evolving state of environmental conditions it may be more appropriate to rely on statements such as "it is highly likely that concentrations of nitrate has been rising over the past decade." If our interest is in the trend that has occurred over some past time period, we can't go back and increase the sample size, and even prospectively, there are limits to the effective sample size that can be achieved even by very frequent sampling, because of the influence of serial correlation. Likelihood statements give decision-makers a clear picture of the degree of certainty that the analyst can give regarding the trends. This is the kind of information they need to take with them as they decide about future actions.

The likelihood of an upwards or downward trend in FNF or FNC can be computed and expressed using the results from the adaptive Bayesian approach described in Section 3.3. For the proposition that FNF is positive, the likelihood, denoted as L_f^+, is determined from the bootstrap replicates as $L_f^+ = (x_f + 0.5)/(M + 1)$, which is the mean of the posterior distribution of π_f after M replicates. Conversely, the likelihood that the true trend in FNF is negative is $L_f^- = 1 - L_f^+$. Similarly, the likelihood that the true trend in FNC is positive is L_c^+, where $L_c^+ = (x_c + 0.5)/(M + 1)$. The likelihood that the true trend in FNC is negative is $L_c^- = 1 - L_c^+$. The EGRETci R-package, which implements the WBT computes these values and also translates them into descriptive statements, in accordance with the definitions given in Table 2.

5.2. Variability of the bootstrap confidence intervals

A bootstrap approach to determining significance levels or confidence intervals will not yield results that are perfectly repeatable because the results are based on the use of random resampling of the data set. Re-running the WBT will result in differences in confidence intervals for trend magnitudes, and may, in some cases result in a different conclusion about rejecting the null hypothesis. This latter problem will typically be restricted to situations in which the computed p-value for the test is close to 0.1. The general concern with the binary nature of the test outcome (reject or fail to reject) that is typical for all NHST procedures, is of even greater concern with a bootstrap procedure because of the potential for successive applications of the test to the same data set to lead to different outcomes. This problem is one of the motivators for using the likelihood statements discussed in Section 4. Expression of the uncertainty results using a likelihood approach treats the test results more as a continuum, ranging from strong indication that there is a trend to a conclusion that the likelihood of an upwards trend is virtually equal to the likelihood of a downwards trend.

The precision with which confidence intervals are determined in the WBT is a function of the number of replicates. The user faces a trade-off between repeatability of results and computational time. Users of the WBT method are free to establish their own comfort level in establishing the degree of repeatability desired. To explore the issue we use as an example the USGS data set for dissolved orthophosphate for the Choptank River, near Greensboro, MD (USGS site 01491000). It consists of 642 sample values over the period of water years 1984–2013. The data are shown in Fig. 5.

The analysis of the concentration record using WRTDS estimates that for the period 2000 through 2013 the FNC increased by 0.0173 mg/L, from 0.0201 mg/L to 0.0374 mg/L, an 86 percent increase over that period. The WRTDS analysis estimates an increase in FNF of 2400 kg/yr (expressed as a change in yield that is 8.4 kg/km^2/yr). This is an increase of 67 percent over the 2000 to 2013 period. Ten repetitions of the WBT were run with $M_{min} = M_{max} = 9$. In each case H_{c0} and H_{f0} were rejected at $a = 0.1$, and in all cases indicating upwards trend in FNC and FNF. Then 10 repetitions were then done with $M_{min} = M_{max} = 40$, and then another 10 with $M_{min} = M_{max} = 100$. Fig. 6 shows the computed confidence intervals in each of these cases. What is clear is that the variability of the end points of the confidence intervals diminishes substantially as the number of replicates increases. It is also clear that FNC confidence intervals are less variable than those for FNF, which is reasonable given the fact that FNF trend results can be strongly influenced by the few observations that were taken at the highest discharges. These six panels provide a qualitative impression of the differing degree of repeatability of the confidence intervals.

From this example, and others considered during the development of the WBT, we can say the following. If a highly reproducable 90 percent confidence interval is considered important to the user of the information, then 9 replicates is clearly insufficient, 40 gives a more consistent result (particularly for FNC), and with 100 replicates the CIs become rather stable. Having said that, it is also clear that the fundamental conclusions in this case do not change as the number of replicates increase. The FNC trend is clear, and even with only 9 replicates it is clear that the trend is positive and in some of the 10 repetitions it appears that the trend is in the range of 60–85 percent and in others may be as high as 70 to 110 percent. Even with 100 replicates (an 11 fold increase in computational time) the results indicate a trend of 60–110 percent in some cases and also results as extreme as about 40–120 percent. The analyst needs to ask the question, how accurately do we feel we need these confidence intervals need to be in order to convey policy relevant information. The WBT provides the computational resources to estimate these confidence intervals, but the determination of the desired level of precision must come from the analyst's understanding of the information user's needs.

5.3. Visualization of uncertainty in WBT through the depiction of a confidence band

One type of output that provides a very helpful visual impression of the trend and the degree of certainty about that trend is to use the WBT methodology to produce a confidence band around the estimated trend. It is conceptually similar to a graphical output from a linear regression analysis, showing the fitted line and a set of confidence intervals around that line. For purposes of demonstration we use the 90% confidence band, but the software allows for other intervals. The interpretation of the 90% confidence band is this: For year j in the record, the confidence band depicts a range of FNF values for which we believe that the probability that FNF_j falls

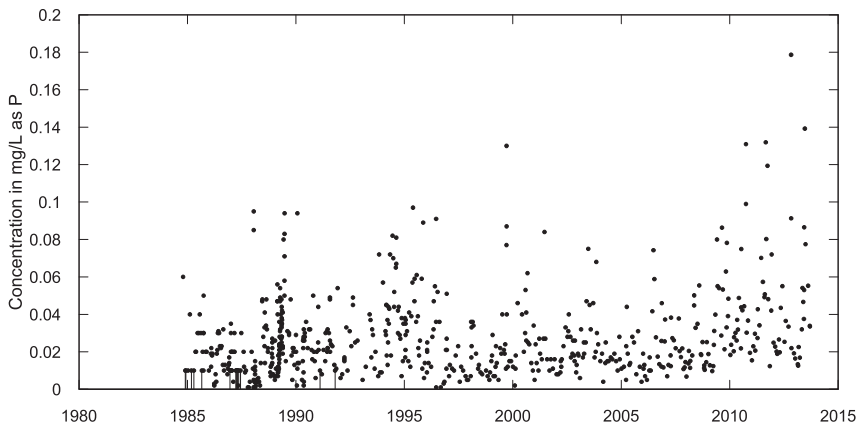

Fig. 5. Concentration of orthosphosphate for the Choptank River near Greensboro, MD, 1984–2013. Vertical lines denote censored values.

inside the band is 90% and a 5% probability that FNF_j lies above it and a 5% probability that it lies below it. The computational process for producing this result is simply to run the WBT process a large number of times (for example, 100 replicates), fit the WRTDS model for every year, and use the fitted model to compute the FNF value for each year for that replicate. Thus, if the number of replicates is 100, then for each year, there are 100 values of FNF computed. The upper and lower bounds for that year are computed by ranking those 100 values and then by interpolating the 5% and 95% quantiles of the sample cumulative distribution function in the same manner as described in section 3.4. Graphically, these limits are shown along with the ordinary WRTDS estimates of the annual FNF. The identical process is carried out for the confidence band for FNC using the same set of bootstrap replicates and the same estimates of *w(Q,T)* that are used for FNF. No formal statistical inferences regarding trend can be made from the resulting graph but the band presents a useful visual means of depicting the "signal" and the "noise" from which to develop a general understanding of whether the changes are large or small in relation to the uncertainty about them.

One feature that is common to these graphs is that at the first few years and last few years the confidence bands can widen

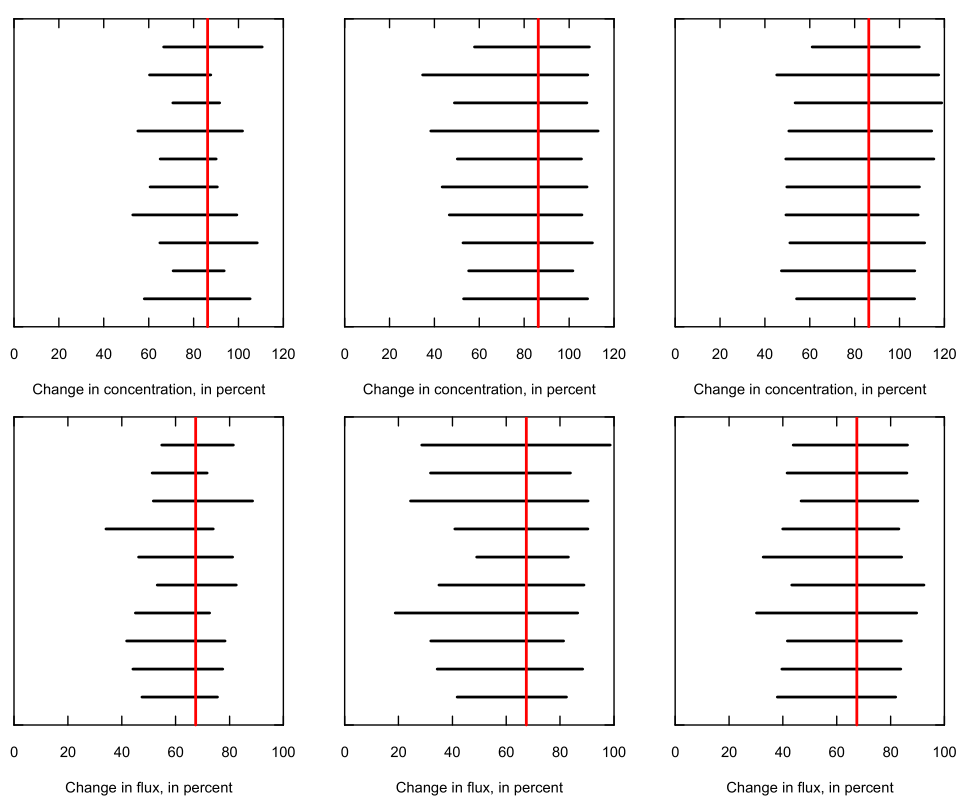

Fig. 6. Repeatability of 90 percent confidence intervals (based on 10 repetitions of the WBT test) for trend in Flow Normalized Concentration (top three panels) or Flow Normalized Flux (bottom three panels), Orthophosphate, Choptank River at Greensboro MD, from 2000 to 2013. Each horizontal black line indicates the WBT estimate of the confidence interval for each of the 10 repetitions of the WBT test. Left panels, show results based on 9 bootstrap replicates, the, center panels are based on 40 replicates and the right panels are based on 100 bootstrap replicates. The vertical red lines indicate the WRTDS estimate of the trend in Flow Normalized Concentration (upper panels) and Flow Normalized Flux (bottom panels) (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

considerably as compared to the middle years. This is exactly what we would expect, because near the ends there is much more uncertainty about the underlying relationships represented by $w(Q,T)$. This same widening is also observed in graphs of confidence intervals around a regression line, but in the case of regression the uncertainty is constrained by the fact that it is assumed that the relationship is linear. Because no such assumption is used in WRTDS, the confidence intervals can diverge to a substantially greater extent than what would be seen in a linear regression model. Also, unlike a linear regression approach, there can be time periods during the record when the width of the bands becomes larger than the years before or after it. This situation is indicative that something happened during that period which increased the inherent variability of the estimates (either a process change or a change in sample density).

The computations needed to produce these confidence bands are much greater than for the other WTB outputs. This is due to two factors. For confidence bands the WRTDS model is estimated for every year in the record on each replicate, but for the other WBT outputs it is only computed for two years, y_s *and* y_e (recall from section 3.2, that the WBT computations do not require bootstrap estimates of $w(Q,T)$ for the entire record but only for these two years, y_s *and* y_e). Thus, for a 30-year record, this can result in a 15-fold increase in computing time over the standard WBT outputs. In addition, the computations for the WBT use the adaptive stopping rule, which is designed specifically for reaching conclusions about two specific hypothesis tests (about FNC and FNF) rather than an overall characterization of uncertainty over all the years of the record. This fact can result in a further increase of 2 or 3 fold in computational time. The EGRETci package provides a script for these computations that employs parallel processing in the R-environment and this greatly speeds up the elapsed time for the computations; however, it can occupy a large part of the computer's processing capacity while it is operating. Because the output is only designed for a visual (and not a numerical) use, the number of repetitions can be fairly small (say <80) and still provide plausibly smooth and reasonably accurate representations of the uncertainty.

The example used here is for total nitrogen for the Susquehanna River at Conowingo, Maryland, USA. The drainage area at this location is 70,200 km^2. This monitoring site is immediately upstream of the Chesapeake Bay. The Susquehanna River is the largest single input of fresh water to Chesapeake Bay, thus these nitrogen inputs are of great importance to ecological conditions in Chesapeake Bay. The data set used covers water years 1985–2013 and consists of 851 samples with a minimum value of 0.557 mg/L, a mean of 1.75 mg/L and a maximum of 20.065 mg/L Fig. 7 shows the confidence band output for FNC and Fig. 8 shows it for FNF.

Both of these figures convey the idea that the overall trend in FNC or FNF has been generally downwards over this period. However, from about 1998 to about 2006, there is an indication of a leveling off or even increasing trend, followed by a continuation of the downward trend in the last seven years of the record. We can see that in both graphs the width of the confidence band was relatively wide at the start and the end of the record, but there was also a period with a wide confidence band from about 2001 to 2006. This appears to be closely related to more extreme variability in the data during that period. Note that many of the individual yearly estimates (shown as the black dots) in these figures lie outside the confidence band. This is not a surprising result. The individual years (particularly for flux) exhibit a great deal of temporal variability due to the variability of discharge. The FN values are computed after removing the year-to-year variability due to discharge. As such they are much less variable. This is again analogous to confidence regions in regression. Many individual observations can lie outside the confidence band.

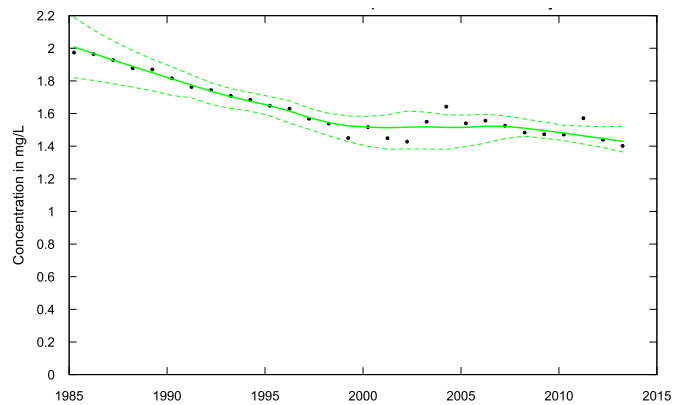

Fig. 7. Susquehanna River at Conowingo, MD results for total nitrogen concentration, showing the 90% confidence band. Calculations use a block length of 200 days and 100 bootstrap replicates. Solid green line shows the annual flow normalized concentrations and the dashed green lines show the 5th and 95th percentiles of the annual flow normalized concentrations. The black dots are the estimated annual mean concentrations (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

Subjectively, looking at these two figures, one can draw a general conclusion that certainly over the full period of 1985–2013 there has been a significant amount of downward trend (both in FNC and FNF). If we just focus on the period 1995 to 2013 we would be inclined to say that the FNC is trending downwards, but for FNF the answer is rather ambiguous, although the indications of downward trend in FNF is moderately clear over the last decade of this period.

5.4. Representation of the distribution of trend magnitudes

Another type of output that can be generated by WBT is a representation of the uncertainty about the trend, expressed as a frequency distribution. Interested parties who wish to evaluate progress towards some established goals might find this representation quite useful. Up to this point in this paper it has simply been assumed that the null hypothesis is that there was no trend. However, managers may be interested in knowing if the improvement was at least as large as some pre-specified goal (e.g. they may

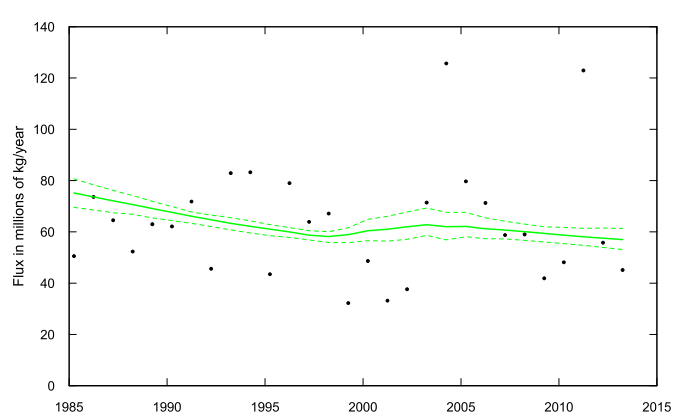

Fig. 8. Susquehanna River at Conowingo, MD, results for total nitrogen flux, showing the 90% confidence band. Calculations use a block length of 200 days and 100 bootstrap replicates. Solid green line shows the annual flow normalized flux and the dashed green lines show the 5th and 95th percentiles of the annual flow normalized flux. The black dots are the estimated annual mean flux (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

have stated that the goal is a 20% reduction in flux over a decade). We can re-express the trends as a percentage change between the starting year and the ending year and show these percentage changes as a histogram. This is accomplished by conducting the WBT (M_{min} should be at least 50 for this application), and saving the computed trend values from each bootstrap replicate, and then re-expressing that trend in terms of a percentage change from the starting year to the ending year (or alternatively they could be shown in actual units such as mg/L or kg/year). Fig. 9 presents histograms for FNC trends and FNF trends as percentage changes from 2000 to 2012 using the same data set presented in Figs. 7 and 8. For these figures M_{min} was set to 200 to create a relatively smooth histogram. Note that both figures indicate strong evidence for a downward trend, but the evidence is slightly weaker in the case of FNC (left panel, likelihood of a downwards trend is 0.78) than for FNF (right panel, likelihood of a downwards trend is 0.89). It is also worth noting that the probability of a large downward trend (say a decrease of more than 10 percent) is much greater for FNF than for FNC even though the standard WRTDS estimate of trend in percent is virtually equal for FNC and FNF (slightly less than a 5 percent decrease).

Consider the following example of a way this information could be used. Parties involved in a Total Maximum Daily Load (TMDL) implementation may establish statements such as this: "our goal is to reduce loadings of nitrogen by 15% from 2000 to 2012." In terms of WRTDS this could be stated as $FNF_{2012} \leq 0.85 \cdot FNF_{2000}$. For this application, one might conduct 100 bootstrap replicates. From each replicate we compute $DR_i = (FNF_{2012i} - FNF_{2000i})/FNF_{2000i}$, the difference between the two years' FNF values expressed as a ratio to the starting year value. Then, the empirical distribution of these 100 values of DR_i can be explored in various ways. The most directly relevant to the goal is this: What is the likelihood that FNF of nitrogen has declined by 15 percent over this time period? The answer would be determined by evaluating the fraction of the $DR_i < 0.85$. In this case it is 4 out of 100 replicates so we could say that the likelihood is 4 percent that the trend from 2000 to 2012 was a decrease of 15 percent or greater (the area to the left of −15 in the right panel histogram). We can also say that the likelihood is 50 percent that the trend over this period was a decrease of 5 percent or more (the area to the left of −5 in the right panel histogram). The software also makes it possible to compute the likelihood that the change between any pair of years is greater than some magnitude (e.g. a decrease of $\leq 6 \cdot 10^6$ *kg/yr*).

6. Summary and conclusions

The WRTDS method was developed as a tool for exploratory data analysis to be used with data sets of water quality sample concentration values and continuous daily discharge values. The method facilitates insights about a number of features of water quality records: including the temporal patterns of trends (including identifying non-monotonic trends), seasonal differences in trend characteristics, difference in the trends for high flow versus low flow, and it provides for an internally consistent method for evaluating the changes in concentration and changes in flux. The original method and associated software (Hirsch and De Cicco, 2014) lacked any capability of describing the uncertainties associated with the WRTDS outputs. This paper describes the theory and implementation of the WRTDS Bootstrap Test (WBT), which enhances WRTDS by adding the capability to quantify the uncertainty of estimated trends and produce graphical representations of the computed uncertainties.

The full set of computations related to uncertainty of trend results is called the WRTDS Bootstrap Test (WBT). Trend is defined by the change between two selected years of a record (y_s *and* y_e). The WBT provides four types of outputs: 1) hypothesis tests for trend in FNC and FNF (reject or do not reject the null hypothesis at $\alpha = 0.1$), 2) p-values for those tests, 3) 90% confidence intervals for the magnitude of the trend in FNC and FNF, 4) likelihood statements (in numerical form and as descriptive statements) about both increasing trends and decreasing trends in FNC and FNF. The R-code for the WBT is available from CRAN. The package is named EGRETci and is dependent upon the EGRET R-package, which houses the R-code for WRTDS analysis.

Using simulated data sets that are realistic statistical models of water quality and discharge time series, the WBT has been shown to result in Type I error rates for a null hypothesis of no trend that are slightly higher than the nominal Type I error rate. It should be noted that other common methods of trend analysis also depart from the nominal Type I error rate by amounts that can be as much or more than the WBT approach. One test, the Seasonal Kendall Test Adjusted for Serial Correlation (SEAKENA) on residuals from the log concentration to log discharge quadratic model, shows performance that is better able to achieve the desired Type I error rate. If

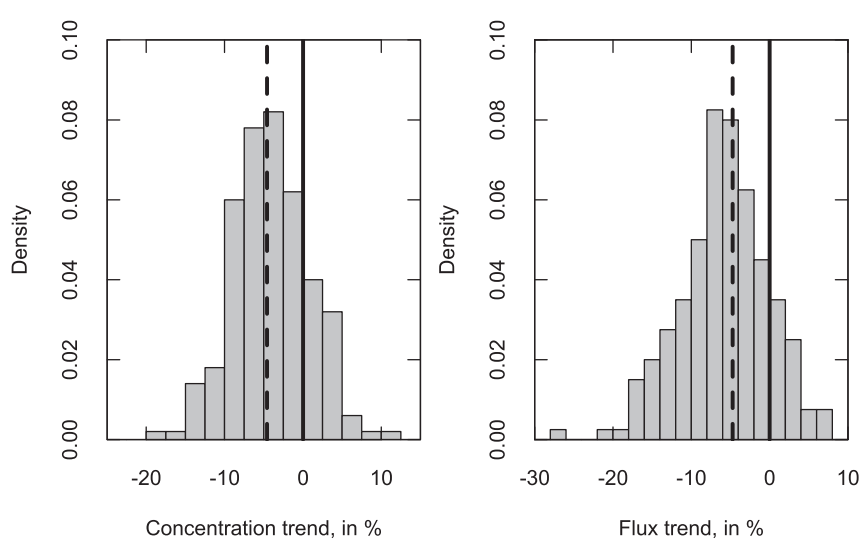

Fig. 9. Histograms of estimated trend magnitudes in percent, 2000–2012, for total nitrogen, Susquehanna River at Conowingo, MD. Left panel is trend in flow-normalized concentration. Right panel is trend in flow-normalized flux. In both cases the standard WRTDS estimate of trend is shown by the dashed vertical line.

the objective of an analysis were solely to report a highly accurate p-value for a hypothesis test for monotonic trend, then SEAKENA would be a good choice. But, if the intent is to have a more complete and flexible method of describing the evolving nature of water quality conditions (both concentration and fluxes) and reasonably accurate metrics of uncertainty, then the WRTDS method in conjunction with WBT would be an appropriate choice. Testing of various stochastic models of streamflow and water quality showed that a bootstrap block length of 200 days is appropriate. The Type I error rate was also found to be insensitive to the level of censoring when up to 50% of the water-quality data was censored.

Analysis of long-term water quality data continues to make progress and future enhancements will come from the development of new methods that combine more deterministic aspects of water quality along with the statistical aspects. Enhancements will also come from new understanding of the temporal properties of water quality time series as new high-frequency data sets (Pellerin et al., 2014) become more common. Water quality statistical analysis has long suffered from the serious challenges in estimating the time-series structure of residuals from statistical models based on discharge, season and time. As longer high frequency data sets become more commonly available these will help to enhance our ability to make more definitive statements about the nature of the non-stationarity of water quality and help us clearly articulate our uncertainties. Ultimately this will facilitate tests better than the one introduced here.

The processes that influence water quality in a river, over periods of multiple decades necessitate a highly flexible representation of relationships of concentration to time, to discharge, and to season. The primary goal of the WRTDS method remains the same as when introduced in 2010: to describe the evolving nature of the changes that are taking place in the system. The addition of the WBT to the WRTDS method provides the added dimension of being able to quantify and communicate to stakeholders the degree of uncertainty that should be attached to findings about trend that emerge from the WRTDS analysis.

Acknowledgments

Author contributions: R.H. designed and performed the research; R.H. and S.A. wrote the paper; R.H. and L.D. wrote the software; L.D. packaged the software. We also acknowledge the assistance with coding by Jeffrey Chanat, advice on time series and hypothesis testing issues from Timothy Cohn, and a very helpful review from Aldo Vecchia. This research has been funded by the U.S. Geological Survey Hydrologic Research and Development Program, National Water Quality Assessment Program, and Chesapeake Bay Ecosystem Program.

Appendices A and B. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.envsoft.2015.07.017.

References

Ames, D.P., 2006. Estimating 7Q10 confidence limits from data: a bootstrap approach. J. Water Resour. Plan. Manage 132, 204–208.
Aulenbach, B.T., Hooper, R.P., 2006. The composite method: an improved method for stream-water solute load estimation. Hydrol. Process 20, 3029–3047.
Boesch, D., Cohn, T.A., Eshleman, K., Grizzard, T.J., Hamlett, J.M., Prestegaard, K.L., Staver, K.W., Weller, D.E., 2005. Assessing Progress and Effectiveness through Monitoring Rivers and Streams.
Broussard III, W.P., Turner, R.E., Westra, J.V., 2012. Do federal farm policies influence surface water quality? Agric. Ecosyst. Environ. 158, 103–109. http://dx.doi.org/10.1016/j.agee.2012.05.022.
Cleveland, W.S., Devlin, S.J., 1988. Locally weighted regression: an approach to regression analysis by local fitting. J. Am. Stat. Assoc. 83, 596–610. http://dx.doi.org/10.1080/01621459.1988.10478639.
Cohn, T.A., Lins, H.F., 2005. Nature's style: naturally trendy. Geophys. Res. Lett. 32, L23402. http://dx.doi.org/10.1029/2005GL024476.
Congdon, P., 2007. Bayesian Statistical Modelling. John Wiley & Sons.
Copeland, C., 2006. Water Quality: Implementing the Clean Water Act. Congressional Research Service, Washington, DC.
Corsi, S.R., De Cicco, L.A., Lutz, M.A., Hirsch, R.M., 2015. River chloride trends in snow-affected urban watersheds: increasing concentrations outpace urban growth rate and are common among all seasons. Sci. Total Environ. 508, 488–497.
Darken, P., Holtzman, G., Smith, E., Zipper, C.E., 2000. Detecting changes in trends in water quality using modified Kendall's tau. Environmetrics 11, 423–434.
Darken, P.F., Zipper, C.E., Holtzman, G.I., Smith, E.P., 2002. Serial correlation in water quality variables: estimation and implications for trend analysis. Water Resour. Res. 38, 22–1.
Davison, A.C., Hinkley, D.V., 1997. Bootstrap Methods and Their Applications, Cambridge Series in Statistical and Probabilistic Mathematics.
Diaconis, P., Efron, B., 1983. Computer-intensive methods in statistics. Sci. Am. 248, 116–130.
Dubrovsky, N.M., Burow, K.R., Clark, G.M., Gronberg, J., Hamilton, P.A., Hitt, K.J., Mueller, D.K., Munn, M.D., Nolan, B.T., Puckett, L.J., 2010. The Quality of Our Nation's Water: Nutrients in the Nation's Streams and Groundwater, 1992-2004. U.S. Geological Survey, Circular. US Department of the Interior, US Geological Survey.
Efron, B., 1979. Bootstrap methods: another look at the jackknife. Ann. Stat. 1–26.
Efron, B., 2005. Bayesians, frequentists, and scientists. J. Am. Stat. Assoc. 100, 1–5.
Efron, B., Tibshirani, R.J., 1994. An Introduction to the Bootstrap. CRC press.
Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2014. Bayesian Data Analysis. Taylor & Francis.
Han, H., Allan, J.D., Bosch, N.S., 2012. Historical pattern of phosphorus loading to Lake Erie watersheds. J. Gt. Lakes Res. 38, 289–298.
Hirsch, R.M., De Cicco, L.A., 2014. User Guide to Exploration and Graphics for RivEr Trends (EGRET) and DataRetrieval: R Packages for Hydrologic Data. U.S. Geological Survey, Reston, VA. U.S. Geological Survey Techniques and Methods 4A-10.
Hirsch, R.M., Slack, J.R., 1984. A nonparametric trend test for seasonal data with serial dependence. Water Resour. Res. 20, 727–732.
Hirsch, R.M., Slack, J.R., Smith, R.A., 1982. Techniques of trend analysis for monthly water quality data. Water Resour. Res. 18, 107–121.
Hirsch, R.M., Moyer, D.L., Archfield, S.A., 2010. Weighted regressions on time, discharge, and season (WRTDS), with an application to Chesapeake Bay river Inputs1. JAWRA J. Am. Water Resour. Assoc. 46, 857–880. http://dx.doi.org/10.1111/j.1752-1688.2010.00482.x.
Ide, J., Chiwa, M., Higashi, N., Maruno, R., Mori, Y., Otsuki, K., 2012. Determining storm sampling requirements for improving precision of annual load estimates of nutrients from a small forested watershed. Environ. Monit. Assess. 184, 4747–4762.
International Joint Commission, 2014. A Balanced Diet for Lake Erie: Reducing Phosphorus Loadings and Harmful Algal Blooms. http://www.ijc.org/files/publications/2014%20IJC%20LEEP%20REPORT.pdf.
Jeffreys, H., 1998. The Theory of Probability. Oxford University Press.
Kirchner, J.W., 2006. Getting the right answers for the right reasons: linking measurements, analyses, and models to advance the science of hydrology. Water Resour. Res. 42.
Kirchner, J.W., Neal, C., 2013. Universal fractal scaling in stream chemistry and its implications for solute transport and water quality trend detection. Proc. Natl. Acad. Sci. 110, 12213–12218. http://dx.doi.org/10.1073/pnas.1304328110.
Knopman, D.S., Smith, R.A., 1993. 20 years of the clean water act has US water quality improved? Environ. Sci. Policy Sustain. Dev. 35, 16–41.
Lahiri, S., Zhu, J., 2006. Resampling methods for spatial regression models under a class of stochastic designs. Ann. Stat. 34, 1774–1813.
Langan, S., Johnston, L., Donaghy, M., Youngson, A., Hay, D., Soulsby, C., 2001. Variation in river water temperatures in an upland stream over a 30-year period. Sci. Total Environ. 265, 195–207.
Langland, M.J., Raffensperger, J.P., Moyer, D.L., Landwehr, J.M., Schwarz, G.E., 2007. Changes in Streamflow and Water Quality in Selected Nontidal Basins in the Chesapeake Bay Watershed, 1985-2004. Government Printing Office. U.S. Geological Survey, Scientific Investigations Report 2006-5178.
Lettenmaier, D.P., 1976. Detection of trends in water quality data from records with dependent observations. Water Resour. Res. 12, 1037–1046.
Mastrandrea, M.D., Mach, K.J., 2011. Treatment of uncertainties in IPCC Assessment Reports: past approaches and considerations for the Fifth Assessment Report. Clim. Change 108, 659–673.
Mehan, T.G., 2012. Adaptation going forward. Water Environ. Technol. 37–38.
Morton, R., Henderson, B.L., 2008. Estimation of nonlinear trends in water quality: an improved approach using generalized additive models. Water Resour. Res. 44.
National Research Council, 2011. Achieving Nutrient and Sediment Reduction Goals in the Chesapeake Bay: an Evaluation of Program Strategies and Implementation. National Academies Press, Washington, DC.
Nicholls, N., 2001. Commentary and analysis: the insignificance of significance testing. Bull. Am. Meteorol. Soc. 82, 981–986.
Pellerin, B.A., Bergamaschi, B.A., Gilliom, R.J., Crawford, C.G., Saraceno, J., Frederick, C.P., Downing, B.D., Murphy, J.C., 2014. Mississippi river nitrate loads

from high frequency sensor measurements and regression-based load estimation. Environ. Sci. Technol. 48, 12612–12619.
Politis, D.N., Romano, J.P., 1994. The stationary bootstrap. J. Am. Stat. Assoc. 89, 1303–1313.
Press, S.J., Press, J.S., 1989. Bayesian Statistics: Principles, Models, and Applications. Wiley, New York.
Rajagopalan, B., Lall, U., 1999. A k-nearest-neighbor simulator for daily precipitation and other weather variables,. Water Resour. Res. 35, 3089–3101.
Reckhow, K.H., Qian, S.S., 1994. Modeling phosphorus trapping in wetlands using generalized additive models. Water Resour. Res. 30, 3105–3114.
Richards, R.P., Baker, D.B., 2002. Trends in water quality in LEASEQ rivers and streams (Northwestern Ohio), 1975–1995. J. Env. Qual. 31, 90–96. http://dx.doi.org/10.2134/jeq2002.9000.
Rustomji, P., Wilkinson, S., 2008. Applying bootstrap resampling to quantify uncertainty in fluvial suspended sediment loads estimated using rating curves. Water Resour. Res. 44.
Ryberg, K.R., Vecchia, A.V., Gilliom, R.J., Martin, J.D., 2014. Pesticide Trends in Major Rivers of the United States, 1992– 2010. United States Geological Survey, Scientific Investigations Report 2014-5135.
Sanborn, A., Hills, T., 2014. The frequentist implications of optional stopping on Bayesian hypothesis tests. Psychon. Bull. Rev. 21, 283–300. http://dx.doi.org/10.3758/s13423-013-0518-9.
Sogbedji, J.M., McIsaac, G.F., 2006. Evaluation of the ADAPT model for simulating nitrogen dynamics in a tile-drained agricultural watershed in central Illinois. J. Environ. Qual. 35, 1914–1923.
Sprague, L., Clark, M., Rus, D., Zelt, R., Flynn, J., Davis, J., 2006. Nutrient and Suspended-sediment Trends in the Missouri River Basin, 1993–2003: US Geological Survey Scientific Investigations Report 2006-5231, 80 p. U.S. Geological Survery.
Stedinger, J., Vogel, R., Foufoula-Georgiou, E., 1993. In: Maidment, D.R. (Ed.), Frequency Analysis of Extreme Events, Chapter 18 in Handbook of Hydrology. McGraw-Hill, New York, 18–1 to 18–66.
Tobin, J., 1958. Estimation of relationships for limited dependent variables. Econom. J. Econom. Soc. 24–36.
van Belle, G., Hughes, J.P., 1984. Nonparametric tests for trend in water quality. Water Resour. Res. 20, 127–136.
Vigiak, O., Bende-Michl, U., 2013. Estimating bootstrap and Bayesian prediction intervals for constituent load rating curves. Water Resour. Res. 49, 8565–8578.
Vogel, R., Rosner, A., Kirshen, P., 2013. Brief Communication: likelihood of societal preparedness for global change: trend detection. Nat. Hazards Earth Syst. Sci. 13, 1773–1778.
Wood, S., 2006. Generalized Additive Models: an Introduction with R. CRC press, Boca Raton, FL.