Response to Editor Comments

Authors' responses inserted as blue text.

As I know both reviewers personally well and consider them as very much competent to judge your work, I decided to thoroughly and carefully review your manuscript to make up my mind in an as much as possible independent manner. As your manuscript is technically demanding, this took a little while. After this effort I am fully convinced that your work is very valuable for the field of hydrology, and yet I think needs moderate revisions before it can be accepted.
Response: Thank you very much for your thoughtful handling of this manuscript and for the additional efforts and time that you invested in as an additional reviewer.

Within this process I encourage you to address the recommendations of Reviewer 2 as outlined in your response.
Response: Thanks. We have thoroughly revised the manuscript based on our prior response to Reviewer 2.

While Reviewer 1 did, although she/he recommended major revisions, did not come up with a single major point that can be addressed, I leave it to you whether you consult a selection of the literature sources he/she provided or not.
Response: We have chosen not to modify the paper after Reviewer 1's comments. As we expressed in our prior responses to Review 1, we believe this is a useful contribution to the HESS community. We are very pleased to learn that both Reviewer 2 and the Editor have seen the merits of the manuscript.

Finally, although I am not an expert in time series analysis, I wonder about the following points:

1. In spatial statistics/geostatistics it is quite common to estimate the spatial covariance in a data set which is irregularly sampled in space by means of the semi-variogram. This implies to pool data into irregularly spaced lag classes and to fit the best suited theoretical variogram function, which of course might be also a power law. It is also standard to fill gaps between the sampling locations by interpolations methods that preserve the covariance structure for instance by means of ordinary kriging or external drift kriging, which may account for a drift in the expectation values (see next comment). I wonder whether the same procedure is not applicable to irregularly sampled time series, both to estimate the autocorrelation function and to fill gaps in a manner that preserves the autocorrelation?

Response: Great comment. While we think it is beyond the scope of our current manuscript, we agree that in future efforts it is useful to develop approaches to estimate autocorrelation (in the time domain) and then use that to fill the gaps in monitoring records. This kind of research can really improve our capability to build statistical models for estimating constituent concentrations and loads for days without samples. Unfortunately, this has been only well understood for simple autocorrelation models such as the autoregressive lag-1 model (AR[1]), which is a short-term memory model. But as we discuss in Section 1.1, water quality data generally exhibit long-term memory. Therefore, it remains a real challenge and also a critical need to "fill gaps in manner that preserves the autocorrelation" for long-memory time series.

Moreover, we want to add that since the auto-covariance is the Fourier transform of the spectrum, estimating the autocorrelation to infill the data before estimating the spectrum is really an example of 'begging the question' (in the original meaning of that idiom) in the sense of assuming a result in order to justify the conclusion.

2. The term water quality data seem a little imprecise. Do you refer to concentrations of solutes in stream flow (C) or to total loads (Q*C) in stream flow. Only the latter reflects mass conservation, and concentration data might exhibit quite a bit of pseudo-dynamics and pseudo seasonality due to simple dilution effects and seasonality in discharge (as you pointed out). In this context there is also quite a bit of discussion in the geostatistical literature whether data sets which include an external drift (a deterministic dependence of their mean on an external variable such as air temperature and or rainfall and topographic elevation) or in your case a flow dependence of concentration shall be detrended before estimating the covariance or not. I personally think it is necessary to remove a drive to avoid mixing of deterministic and stochastic variability – and there is evidence that a variogram analysis without the removal of a trend leads to a positive bias in the range and the sill of the variogram (e.g. Zimmermann et al. 2008).

Response: Thank you for this note on "external drift." We agree and that is why we discussed the complexities in concentration data with respect to trend, discharge, and season effects in the Introduction (see lines 139-153 in Section 1.3) and why we applied the WRTDS method to the real concentration data to remove such effects prior to the fractal scaling estimation (see Section 3.3).

3. The definition of delta t average seems inconsistent with example 2 in line 199.

Response: We double checked the definition and found the example consistent with its definition. For example 2, delta t average = (2+1+4+3)/4 = 2.5 hr, which also equals to L/(N-1) = 10/(5-1) = 2.5 hr.

4. Please flip the numbering of figure 1 and 2.

Response: Please note that Figure 1 appears first – below equation (8) in Section 1.2, so it is not needed to flip the two figures.

Response to Anonymous Referee #2

Authors' responses inserted as blue text.

I find the manuscript is well-written and technically rigorous, with results that can be generalized beyond hydrologic time series. This manuscript tackles a challenging and highly relevant topic - the quantification of fractal scaling behavior for irregularly sampled data - and provides needed synthesis on the most promising methods to estimate this behavior. For these reasons, I recommend the manuscript be accepted subject to minor revisions.

Response: Thank you for these comments.

I do, however, have a number of comments that would help improve clarity of the manuscript and emphasize the more practical aspects of this work.

Major comments:

1) Lines 127-129: It would be interesting to the reader and for understanding the important contribution of this work to detail the effects of non-normal data and persistence, seasonality, and the presence of long-term trends on the estimation of Beta.

Response: We agree this would be potentially useful to do, but it is simply beyond the scope of the paper -- new modeling experiments for each of these effects would multiply the length and complexity of the paper by large factors. We recognize this would be an important area for future research, which we explicitly put in the Section 3.3 (lines 459-464): "*such real data are typically much more complex than our simulated time series, because of (1) strong deviations from normal distributions and (2) effects of flow-dependence, seasonality, and temporal trend (Hirsch et al., 1991; Helsel and Hirsch, 2002). In this regard, future research may simulate time series with these important characteristics and evaluate the performance of various estimation approaches, perhaps following the modeling framework described herein.*"

2) Lines 264-265: It is noted that the results which demonstrate that the approach used in this manuscript to mimic the sampling irregularity performs well as compared to other simulation methods are not shown. I think these results are important to show, as this approach is what underlies the remainder of the analysis of the methods. This can be added to the supplementary material.

Response: Thank you for this suggestion. We now provide these simulation results in the supplementary material, see *Section S1*.

3) There are a large number of interpolation methods (n=11) presented here. I would argue that some of these methods are not very realistic in the context of what one would experience in terms interpolation for irregular samples. Unless the authors provide sound technical justification for each scenario, I would consider removing scenarios that would not generally be considered in standard

practices (examples are scenarios B3, B4, and select a smaller subset of LOESS smoothing parameter values). This would also streamline the results and text.
Response: The interpolation methods were selected not only on the basis of their frequency of use, but to ensure a certain degree of completeness -- we felt it would ultimately be more useful to include obvious variations of common methods than to exclude one that someone might have been considering and looked to our paper for guidance. Although methods B3-B5 seem not plausible, they have been discussed and used in the scientific literature (e.g., Blankers et al. 2010; Graham 2009). Furthermore, some R packages have been developed (e.g., "na.locf"), making these methods readily available to general users without any prior knowledge on the methods' performance. Therefore, we think it is worthwhile to keep the results and discussion of these methods' performance, which is a useful contribution to the topic of "missing data analysis." Furthermore, removing one or two methods would do little to shorten or simplify the paper at this stage, so we have chosen not to so.

4) Line 412: For Monte Carlo analysis, average values of the simulated parameter of interest are computed from sample sizes of 100 or more - not 30. Was this tested in your experiments?
Response: We used 30 samples because it was quite sufficient to constrain estimates of the average in most cases. Standard error of the mean beta for most methods is much smaller than the variation between methods. Nonetheless, to follow common practice, we have adopted this suggestions to run the simulation 100 times and we have revised all relevant figures and text accordingly (including Figures 5-9 and Figures S3-S12.

Minor clarification comments:

1) Lines 3-4: Consider adding a phrase or sentence to explain why spectral slope is important to trend detection.
Response: We agree and have added the phrase "to avoid false inference on the statistical significance of trends" at the end of this sentence.

2) Line 15: Is the "modified form" being newly introduced here? Or does it already exist. Clarify.
Response: It is a published method, see method C1b in Section 3.1 where it is introduced. For clarity, we have revised "a modified form" to "a previously-published modified form."

3) line 38-39: The fact that ACF is summable seems a non-sequitor here. It is later that the connection is made to summability and the presence of fractal behavior. Perhaps it is not necessary to comment on the summability of the ACF?
Response: Agreed. We have deleted the comment on "summability".

4) lines 90-103: Moving this paragraph to the end of Section 1.1 would provide more immediate clarity as to the scope and value of this work.
Response: Thanks. This paragraph has been moved to the suggested location, i.e., Section 1.1.

5) Line 158: Consider the of the work “interpolating” instead of “modeling”
Response: We have decided to keep the use of “modeling.”

6) Section 2.1: Highly clever way to define sampling irregularity.
Response: Thank you.

7) line 216 (and throughout): I do not think “gappy” is a word and chuckled at its appearance. Please replace with “irregularly-spaced”.
Response: Thank you. We have replaced the word “gappy” with “irregularly-spaced” throughout the manuscript.

8) line 237: Please be more specific in how you arrived at this equation for the shape parameter.
Response: As the line states, $\lambda = \mu^2/[\mathrm{var}(\Delta t^*) - \mu] = (\mathrm{mean}(\Delta t^*) - 1)^2/[\mathrm{var}(\Delta t^*) - \mathrm{mean}(\Delta t^*) + 1]$. The first equality represents a well-established property for the negative binomial distribution. The second equality is achieved through the substitution of μ by “$\mathrm{mean}(\Delta t^*) - 1$”. We think it is already clear and hence further modification is not necessary.

9) line 247 (as an example): Please add units to values provided in this section and throughout. This will help the reader follow the results and methods.
Response: Units are provided for $\Delta t_{average}$ (days). Note that λ and μ are fitted (negative binomial distribution) parameters for the *non-dimensionalized* time series (Δt^*) -- see Section 2.1 -- therefore these two variables do not have units.

10) line 473: Hirsch and DeCicco (2015) is the reference to the user manual for WRTDS. The method itself is explained in Hirsch et al. (2010). I would cite the original paper.
Hirsch, R. M., Moyer, D. L. and Archfield, S. A. (2010), Weighted Regressions on Time, Discharge, and Season (WRTDS), with an Application to Chesapeake Bay River Inputs. JAWRA Journal of the American Water Resources Association, 46: 857–880. doi: 10.1111/j.1752-1688.2010.00482.x
Response: Thanks. We have corrected this citation to Hirsch et al. (2010).

References Cited

Blankers, M., Koeter, M. W., & Schippers, G. M. (2010). Missing data approaches in eHealth research: simulation study and a tutorial for nonmathematically inclined researchers. Journal of medical Internet research, 12(5).

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. Annual review of psychology, 60, 549-576.

Response to Anonymous Referee #1

Authors' responses inserted as blue text.

The study performs a comparative evaluation of statistical methods available in the literature using as benchmark test the quantification of fractal scaling in water quality time series with irregular sampling. While lacking technical novelty, the manuscript is written in a sober, careful manner, aptly guiding the reader throughout the key arguments, methodologies and procedures, and presenting the results in a clear and concise manner.

Response: While we thank the reader for the comments about the clarity of our presentation, and for the helpful comments the reviewer has made, we disagree with the comment about the lack of technical novelty, which is repeated elsewhere below. We respond to it there.

This being said, I raise four fundamental concerns:

1) The paper is a pure statistical exercise. It would thus benefit from a physical interpretation of the methodological structure and results: namely, discussing possible physical mechanisms responsible for the statistical signatures detected in the analysis, along with their physical consistency. For instance, whether a trend is physically sustainable and realistic in terms of system energetics, what physical mechanisms sustain the power laws detected in the data, and what physically entails the fractal behaviour. Fractals and scaling are well understood in the physical sciences but the HESS readers will be happy to learn this in the hydrological context. In doing so, the authors will be able to strengthen their arguments and diffuse concerns about whether there is any realism underlying the signatures detected in the analysis.

Response: While we appreciate the value of a physical interpretation, we will not alter the manuscript in response to this comment for the following reasons:

I. We disagree that this is a 'statistical exercise' (which we take to imply that it provides no actual illumination). As its title reflects, the manuscript focuses on the comparison of various statistical methods for quantifying fractal scaling. These methods are actually used in the literature, and we believe that it is of value to know if they actually work or not.

II. Sections 1.1 and 1.2 in the Introduction provide background information with some reference to the physical origins, and providing many citations to literature that provides the insights the reviewer is looking for. In fact, significant sections of the foundational literature on fractals and scaling were developed within the hydrology community and later picked up more widely (the Hurst exponent is an example).

III. The physical interpretations of fractal time series are many and varied, depending on the context. Those interpretations are already available in the literature, and our purpose is neither to add to them, nor to review them. Instead our purpose is to conduct benchmark tests to determine whether some widely used techniques for inferring fractal scaling are reliable or not.

2) Monte Carlo simulations can be structured and tuned for essentially any purpose and to yield any outcome, relying on wise choices made in the methodological setup and the generating system, based on the researchers' understanding or conception of its behaviour. If the methodological setup is entirely data-based, i.e. learn from some statistic or machine learning procedure derived from dataset records, there will always be a degree of case-specific empiricism that is not straightforwardly generalisable, unless there is a fundamental principle beneath. This again links to concern 1. Therefore, it is important to thoroughly provide a solid background to all the assumptions supporting the choices made in the methodological setup and operation.
Response: We believe we have provided sufficient reasoning for the assumptions adopted, but welcome comments on any specific area that is unclear or has been omitted.

3) The paper does not introduce any methodological novelty. In fact, there is a vast literature on statistics of irregularly sampled series (also known as unevenly spaced time series). Therefore, I strongly encourage the authors to look into the literature outside of hydrology, e.g. in astrophysics, neurosciences, paleoclimatology, where they will find a rich diversity of sophisticated and long-proven methods that already tackle the same problems. In doing so, the authors will necessarily tone down the false claims about novelty in new methods and frameworks, when in reality the only novelty is the application of existing methods to hydrological case studies. The key merit of the paper is essentially the comparative evaluation of well known statistical methods and their application to the hydrological sciences, namely relevant water quality issues. As such, this is a purely applied paper and should be clearly presented as such. This brings me to the fourth concern.
Response: We agree with the reviewer's comment that "the paper is essentially the comparative evaluation of well known statistical methods and their application to the hydrological sciences, namely relevant water quality issues." This is precisely stated in Section 1.3 where we define the scope of the work -- see the 2nd last paragraph in that section. But we stress that we never claimed that our work is about developing "new methods". Rather, it is stated clearly in several locations of the paper that this work is about the evaluation of existing statistical methods.

In addition, we disagree with the reviewer on the point that there are "a rich diversity of sophisticated and long-proven methods that already tackle the same problems". Many existing methods do not apply to irregularly sampled data and hence can not be used. Others have been widely used, but have not been rigorously tested. The Lomb-Scargle spectral method is well established, but has known weaknesses (as discussed in the paper and elsewhere, see Montanari et al. 1999). If the reviewer is aware of other works that solve the problem addressed by our work, we again encourage the reviewer to provide citations.

Regarding the novelty (or contribution) of this work, we are not aware of any other papers that perform a similar comparative analysis of these methods, let alone one that is tailored to the needs of the hydrology and earth science community. The reviewer provides no evidence or citations to back up the claim that our study is not novel. If the reviewer is aware of any other such studies we would encourage the reviewer to provide them. One of us has been working on this problem for over 20 years, and in that time has seen no published work that is similar to ours.

Studies that review, compare and critically evaluate available methods are valuable contributions to the scientific literature. They are, in our opinion, useful checks on a proliferation of divergent methods that threatens to generate (at best) incomparable and (at worst) inaccurate observations of physical phenomena.

Our contribution in this regard is explicitly summarized in the last paragraph of the paper, which is copied below: "*Overall, these results provide new contributions in terms of better understanding and quantification of the proposed methods' performances for estimating the strength of fractal scaling in irregularly sampled water-quality data. In addition, the work has provided an innovative and general approach for modeling sampling irregularity in water-quality records. Moreover, this work has proposed and demonstrated a generalizable framework for data simulation (with gaps) and β estimation, which can be readily applied toward the evaluation of other methods that are not covered in this work. More generally, the findings and approaches may also be broadly applicable to irregularly sampled data in other scientific disciplines. Last but not least, we note that accurate quantification of fractal scaling in irregular water-quality time series remains an unresolved challenge for the hydrologic community and for many other disciplines that must grapple with irregular sampling.*"

4) There are no novel hydrological insights in the paper. While the statistical messages are useful (albeit not technically novel), it would be essential to bring out a substantial advance in the understanding of the hydrological and earth systems. After all, HESS is not merely a journal of applied statistics but rather one in which there should be something to be learnt in the functioning of the hydrological system.
Response: We disagree with the reviewer that this work does not provide contributions to HESS in terms of understanding of the hydrological and earth systems. As noted in Section 1.3 "Motivations and Objectives of this Work," the quantification of fractal scaling has important implications for detecting trends in water quality time series, but there is a large gap with respect to what methods are appropriate (or applicable) for quantifying fractal scaling in irregularly sampled water quality time series. By dealing with this issue, this work is highly relevant to the hydrological community.

<u>References Cited</u>

Montanari, A., M. S. Taqqu and V. Teverovsky, 1999. Estimating long-range dependence in the presence of periodicity: An empirical study. *Mathematical and Computer Modelling* 29:217-228, DOI: 10.1016/S0895-7177(99)00104-1.

Evaluation of statistical methods for quantifying fractal scaling in water quality time series with irregular sampling

Qian Zhang[1], Ciaran J. Harman[2], James W. Kirchner[3,4,5]

[1] University of Maryland Center for Environmental Science at the US Environmental Protection Agency Chesapeake Bay Program Office, 410 Severn Avenue, Suite 112, Annapolis, MD 21403 (formerly, Department of Geography and Environmental Engineering, Johns Hopkins University, 3400 North Charles Street, Baltimore, Maryland 21218)

[2] Department of Environmental Health and Engineering, Johns Hopkins University, 3400 North Charles Street, Baltimore, Maryland 21218

[3] Department of Environmental System Sciences, ETH Zurich, Universitätstrasse 16, CH-8092 Zurich, Switzerland

[4] Swiss Federal Research Institute WSL, Zürcherstrasse 111, CH-8903 Birmensdorf, Switzerland

[5] Department of Earth and Planetary Science, University of California, Berkeley, California 94720

Correspondence to: Qian Zhang (qzhang@chesapeakebay.net)

Abstract. River water-quality time series often exhibit fractal scaling, which here refers to autocorrelation that decays as a power law over some range of scales. Fractal scaling presents challenges to the identification of deterministic trends to avoid false inference on the statistical significance of trends, but traditional methods for estimating spectral slope (β) or other equivalent scaling parameters (*e.g.*, Hurst exponent) are generally inapplicable to irregularly sampled data. Here we consider two types of estimation approaches for irregularly sampled data and evaluate their performance using synthetic time series. These time series were generated such that (1) they exhibit a wide range of prescribed fractal scaling behaviors, ranging from white noise (β = 0) to Brown noise (β = 2), and (2) their sampling gap intervals mimic the sampling irregularity (as quantified by both the skewness and mean of gap-interval lengths) in real water-quality data. The results suggest that none of the existing methods fully account for the effects of sampling irregularity on β estimation. First, the results illustrate the danger of using interpolation for gap filling when examining auto-correlation, as the interpolation methods consistently under-estimate or over-estimate β under a wide range of prescribed β values and gap distributions. Second, the long-established Lomb-Scargle spectral method also consistently

under-estimates β. A previously-published modified form, using only the lowest 5% of the frequencies for spectral slope estimation, has very poor precision, although the overall bias is small. Third, a recent wavelet-based method, coupled with an aliasing filter, generally has the smallest bias and root-mean-squared error among all methods for a wide range of prescribed β values and gap distributions. The aliasing method, however, does not itself account for sampling irregularity, and this introduces some bias in the result. Nonetheless, the wavelet method is recommended for estimating β in irregular time series until improved methods are developed. Finally, all methods' performances depend strongly on the sampling irregularity, highlighting that the accuracy and precision of each method are data-specific. Accurately quantifying the strength of fractal scaling in irregular water-quality time series remains an unresolved challenge for the hydrologic community and for other disciplines that must grapple with irregular sampling.

Key Words

Fractal scaling, autocorrelation, Hurst effect, river water-quality sampling, sampling irregularity, trend analysis

1. Introduction

1.1. Autocorrelations in Time Series

It is well known that time series from natural systems often exhibit auto-correlation, that is, observations at each time step are correlated with observations one or more time steps in the past. This property is usually characterized by the autocorrelation function (ACF), which is defined as follows for a process X_t at lag *k*:

$$\gamma(k) = cov(X_t, X_{t+k}) \tag{1}$$

In practice, auto-correlation has been frequently modeled with classical techniques such as auto-regressive (AR) or auto-regressive moving-average (ARMA) models (Darken *et al.*, 2002; Yue *et al.*, 2002; Box *et al.*, 2008). These models assume that the underlying process has short-term memory, *i.e.*, the ACF decays exponentially with lag *k*~~, which implies that the ACF is summable~~ (Box *et al.*, 2008).

Although the short-term memory assumption holds sometimes, it cannot adequately describe many time series whose ACFs decay as a power law (thus much slower than exponentially) and may not reach zero even for large lags, which implies that the ACF is non-summable. This property is commonly referred to as long-term memory or fractal scaling, as opposed to short-term memory (Beran, 2010).

Fractal scaling has been increasingly recognized in studies of hydrological time series, particularly for the common task of trend identification. Such hydrological series include riverflow (Montanari *et al.*, 2000; Khaliq *et al.*, 2008; Khaliq *et al.*, 2009; Ehsanzadeh and Adamowski, 2010), air and sea temperature (Fatichi *et al.*, 2009; Lennartz and Bunde, 2009; Franzke, 2012b; Franzke, 2012a), conservative tracers (Kirchner *et al.*, 2000; Kirchner *et al.*, 2001; Godsey *et al.*, 2010), and non-conservative chemical constituents (Kirchner and Neal, 2013; Aubert *et al.*, 2014). Because for fractal scaling processes the variance of the sample mean converges to zero much slower than the rate of n^{-1} (n: sample size), the fractal scaling property must be taken into account to avoid "false positives" (Type I errors) when inferring the statistical significance of trends (Cohn and Lins, 2005; Fatichi *et al.*, 2009; Ehsanzadeh and Adamowski, 2010; Franzke, 2012a). Unfortunately, as stressed by Cohn and Lins (2005), it is "surprising that nearly every assessment of trend significance in geophysical variables published during the past few decades has failed [to do so]", and a similar tendency is evident in the decade following that statement as well.

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

1.2. Overview of Approaches for Quantification of Fractal Scaling

Several equivalent metrics can be used to quantify fractal scaling. Here we provide a review of the definitions of such processes and several typical modeling approaches, including both time-domain and frequency-domain techniques, with special attention to their reconciliation. For a more comprehensive review, readers are referred to Beran *et al.* (2013), Boutahar *et al.* (2007), and Witt and Malamud (2013).

Strictly speaking, X_t is called a stationary long-memory process if the condition

$$\lim_{k\to\infty} k^{\alpha}\gamma(k) = C_1 > 0 \qquad (2)$$

where C_1 is a constant, is satisfied by some $\alpha \in (0,1)$ (Boutahar *et al.*, 2007; Beran *et al.*, 2013). Equivalently, X_t is a long-memory process if, in the spectral domain, the condition

$$\lim_{\omega\to 0} |\omega|^{\beta} f(\omega) = C_2 > 0 \qquad (3)$$

is satisfied by some $\beta \in (0,1)$, where C_2 is a constant and $f(\omega)$ is the spectral density function of X_t, which is related to ACF as follows (which is also known as the Wiener-Khinchin theorem):

$$f(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma(k) e^{-ik\omega} \tag{4}$$

where ω is angular frequency (Boutahar *et al.*, 2007).

One popular model for describing long-memory processes is the so-called fractional auto-regressive integrated moving-average model, or ARFIMA (p, q, d), which is an extension of ARMA models and is defined as follows:

$$(1 - B)^d \varphi(B) X_t = \psi(B) \varepsilon_t \tag{5}$$

where ε_t is a series of independent, identically distributed Gaussian random numbers ~ (0, σ_ε^2), B is the backshift operator (*i.e.*, $BX_t = X_{t-1}$), and functions $\varphi(\cdot)$ and $\psi(\cdot)$ are polynomials of order p and q, respectively. The fractional differencing parameter d is related to the parameter α in Eq. (2) as follows:

$$d = \frac{1 - \alpha}{2} \in (-0.5, 0.5) \tag{6}$$

(Beran *et al.*, 2013; Witt and Malamud, 2013).

In addition to a slowly decaying ACF, a long-memory process manifests itself in two other equivalent fashions. One is the so-called "Hurst effect", which states that, on a log-log scale, the range of variability of a process changes linearly with the length of time period under consideration. This power-law slope is often referred to as the "Hurst exponent" or "Hurst coefficient" H (Hurst, 1951), which is related to d as follows:

$$H = d + 0.5 \tag{7}$$

(Beran *et al.*, 2013; Witt and Malamud, 2013). The second equivalent description of long-memory processes, this time from a frequency-domain perspective, is "fractal scaling", which describes a power-law decrease in spectral power with increasing frequency, yielding power spectra that are linear on log-log axes (Lomb, 1976; Scargle, 1982; Kirchner, 2005). Mathematically, this inverse proportionality can be expressed as:

$$f(\omega) = C_3 |\omega|^{-\beta} \tag{8}$$

where C_3 is a constant and the scaling exponent β is termed the "spectral slope." In particular, for spectral slopes of zero, one, and two, the underlying processes are termed as "white", "pink" (or "flicker"), and "Brown" (or "red") noises, respectively (Witt and Malamud, 2013). Illustrative examples of these three noises are shown in **Figure 1a-1c**.

In addition, it can be shown that the spectral density function for ARFIMA (p,d,q) is

$$f(\omega) = \frac{\sigma_\varepsilon^2}{2\pi} \frac{\left|\psi(e^{-i\omega})\right|^2}{|\varphi(e^{-i\omega})|^2} \left|1 - e^{-i\omega}\right|^{-2d} \tag{9}$$

for $-\pi < \omega < \pi$ (Boutahar *et al.*, 2007; Beran *et al.*, 2013). For $|\omega| \ll 1$, Eq. (9) can be approximated by:

$$f(\omega) = C_4 |\omega|^{-2d} \tag{10}$$

with

$$C_4 = \frac{\sigma_\varepsilon^2}{2\pi} \frac{|\psi(1)|^2}{|\varphi(1)|^2} \tag{11}$$

Eq. (10) thus exhibits the asymptotic behavior required for a long-memory process given by Eq. (3). In addition, a comparison of Eq. (10) and (8) reveals that,

$$\beta = 2d \tag{12}$$

Overall, these derivations indicate that these different types of scaling parameters (*i.e.*, α, d, and H and β) can be used equivalently to describe the strength of fractal scaling. Specifically, their equivalency can be summarized as follows:

$$\beta = 2d = 1 - \alpha = 2H - 1 \tag{13}$$

It should be noted, however, that the parameters d, α, and H are only applicable over a fixed range of fractal scaling, which is equivalent to (-1, 1) in terms of β.

~~Fractal scaling has been increasingly recognized in studies of hydrological time series, particularly for the common task of trend identification. Such hydrological series include riverflow (Montanari *et al.*, 2000; Khaliq *et al.*, 2008; Khaliq *et al.*, 2009; Ehsanzadeh and Adamowski, 2010), air and sea temperature (Fatichi *et al.*, 2009; Lennartz and Bunde, 2009; Franzke, 2012b; Franzke, 2012a), conservative tracers (Kirchner *et al.*, 2000; Kirchner *et al.*, 2001; Godsey *et al.*, 2010), and non-conservative chemical constituents (Kirchner and Neal, 2013; Aubert *et al.*, 2014). Because for fractal scaling processes the variance of the sample mean converges to zero much slower than the rate of n^{-1} (n: sample size), the fractal scaling property must be taken into account to avoid "false positives" (Type I errors) when inferring the statistical significance of trends (Cohn and Lins, 2005; Fatichi *et al.*, 2009; Ehsanzadeh and Adamowski, 2010; Franzke, 2012a). Unfortunately, as stressed by Cohn and Lins (2005), it is "surprising that nearly every assessment of trend significance in geophysical variables published during the past few decades has failed [to do so]", and a similar tendency is evident in the decade following that statement as well.~~

Formatted: Indent: First line: 0"

1.3. Motivation and Objective of this Work

To account for fractal scaling in trend analysis, one must be able to first quantify the strength of fractal scaling for a given time series. Numerous estimation methods have been developed for this purpose, including Hurst rescaled range analysis, Higuchi's method, Geweke and Porter-Hudak's method, Whittle's maximum likelihood estimator, detrended fluctuation analysis, and others (Taqqu *et al.*, 1995; Montanari *et al.*, 1997; Montanari *et al.*, 1999; Rea *et al.*, 2009; Stroe-Kunold *et al.*, 2009). For brevity, these methods are not elaborated here; readers are referred to Beran (2010) and Witt and Malamud (2013) for details. While these estimation methods have been extensively adopted, they are unfortunately only applicable to regular (*i.e.*, evenly spaced) data, *e.g.*, daily streamflow discharge, monthly temperature, *etc*. In practice, many types of hydrological data, including river water-quality data, are often sampled irregularly or have missing values, and hence their strengths of fractal scaling cannot be readily estimated with the above traditional estimation methods.

Thus, estimation of fractal scaling in irregularly sampled data is an important challenge for hydrologists and practitioners. Many data analysts may be tempted to interpolate the time series to make it regular and hence analyzable (Graham, 2009). Although technically convenient, interpolation can be problematic if it distorts the series' autocorrelation structure (Kirchner and Weil, 1998). In this regard, it is important to evaluate various types of interpolation methods using carefully designed benchmark tests and to identify the scenarios under which the interpolated data can yield reliable (or, alternatively, biased) estimates of spectral slope.

Moreover, quantification of fractal scaling in real-world water-quality data is subject to several common complexities. First, water-quality data are rarely normally distributed; instead, they are typically characterized by log-normal or other skewed distributions (Hirsch *et al.*, 1991; Helsel and Hirsch, 2002), with potential consequences for β estimation. Moreover, water-quality data also tend to exhibit long-term trends, seasonality, and flow-dependence (Hirsch *et al.*, 1991; Helsel and Hirsch, 2002), which can also affect the accuracy of β estimate. Thus, it may be more plausible to quantify β in transformed time series after accounting for the seasonal patterns and discharge-driven variations in the original time series, which is also the approach taken in this work. For the trend aspect, however, it remains a puzzle whether the data set should be de-trended before conducting β estimation. Such de-trending treatment can certainly affect the estimated value of β and hence the validity of (or confidence in) any inference made regarding the statistical significance of temporal trends in the time series. This somewhat circular issue is beyond the scope of our current work -- it has been previously discussed in the context of short-term memory (Zetterqvist, 1991; Darken *et al.*, 2002; Yue *et al.*, 2002; Noguchi *et al.*, 2011; Clarke, 2013; Sang *et al.*, 2014), but it is not well understood in the context of fractal scaling (or long-term memory) and hence presents an important area for future research.

In the above context, the main objective of this work was to use Monte Carlo simulation to systematically evaluate and compare two broad types of approaches for estimating the strength of fractal scaling (*i.e.*, spectral slope β) in irregularly sampled river water-quality time series. Specific aims of this work include the following:

(1) To examine the sampling irregularity of typical river water-quality monitoring data and to simulate time series that contain such irregularity; and

(2) To evaluate two broad types of approaches for estimating β in simulated irregularly sampled time series.

The first type of approach includes several forms of interpolation techniques for gap filling, thus making the data regular and analyzable by traditional estimation methods. The second type of approach includes the well-known Lomb-Scargle periodogram (Lomb, 1976; Scargle, 1982) and a recently developed wavelet method combined with a spectral aliasing filter (Kirchner and Neal, 2013). The latter two methods can be directly applied to irregularly spaced data; here we aim to compare them with the interpolation techniques. Details of these various approaches are provided in **Section 3.1**.

This work was designed to make several specific contributions. First, it uses benchmark tests to quantify the performance of a wide range of methods for estimating fractal scaling in irregularly sampled water-quality data. Second, it proposes an innovative and general approach for modeling sampling irregularity in water-quality records. Third, while this work was not intended to compare all published estimation methods for fractal scaling, it does provide and demonstrate a generalizable framework for data simulation (with gaps) and β estimation, which can be readily applied toward the evaluation of other methods that are not covered here. Last but not least, while this work was intended to help hydrologists and practitioners understand the performance of various approaches for water-quality time series, the findings and approaches may be broadly applicable to irregularly sampled data in many other scientific disciplines.

The rest of the paper is organized as follows. We propose a general approach for modeling sampling irregularity in typical river water-quality data and discuss our approach for simulating irregularly sampled data (**Section 2**). We then introduce the various methods for estimating fractal scaling in irregular time series and compare their estimation performance (**Section 3**). We close with a discussion of the results and implications (**Section 4**).

2. Quantification of Sampling Irregularity in River Water-Quality Data

2.1. Modeling of Sampling Irregularity

River water-quality data are often sampled irregularly. In some cases, samples are taken more frequently during particular periods of interest, such as high flows or drought periods; here we will address the implications of the irregularity, but not the (intentional) bias, inherent in such a sampling strategy. In other cases, the sampling is planned with a fixed sampling interval (*e.g.*, 1 day) but samples are missed (or lost, or fail quality-control checks) at some time steps during implementation. In still other cases, the sampling is intrinsically irregular because, for example, one cannot measure the chemistry of rainfall on rainless days or the chemistry of a stream that has dried up. Theoretically, any deviation from fixed-interval sampling can affect the subsequent analysis of the time series.

To quantify the sampling irregularity, we propose a simple and general approach that can be applied to any time series of monitoring data. Specifically, for a given time series with N points, the time intervals between adjacent samples are calculated; these intervals themselves make up a time series of N-1 points that we call Δt. In addition, the following parameters are calculated to quantify its sampling irregularity:

- L = the length of the period of record,
- N = the number of samples in the record,
- $\Delta t_{nominal}$ = the nominal sampling interval under regular sampling (*e.g.*, $\Delta t_{nominal}$ = 1 day for daily samples),
- $\Delta t^* = \Delta t / \Delta t_{nominal}$, the sample intervals non-dimensionalized by the nominal sampling interval,
- $\Delta t_{average} = L/(N-1)$ the average of all the entries in Δt.

The quantification is illustrated with two simple examples. The first example contains data sampled every hour from 1:00 am to 11:00 am on one day. In this case, L = 10 hours, N = 11 samples, $\Delta t = \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1\}$ hour, and $\Delta t_{nominal} = \Delta t_{average} = 1$ hour. The second example contains data sampled at 1:00 am, 3:00 am, 4:00 am, 8:00 am, and 11:00 am. In this case, $L = 10$ hours, $N = 5$ samples, $\Delta t = \{2, 1, 4, 3\}$ hours, $\Delta t_{nominal} = 1$ hour, and $\Delta t_{average} = 2.5$ hours. It is readily evident that the first case corresponds to fixed-interval (regular) sampling that has the property of $\Delta t_{average}/\Delta t_{nominal} = 1$ (dimensionless), whereas the second case corresponds to irregular sampling for which $\Delta t_{average}/\Delta t_{nominal} > 1$.

The dimensionless set Δt^* contains essential information for determining sampling irregularity. This set is modeled as independent, identically distributed values drawn from a negative binomial (NB) distribution. This distribution has two dimensionless parameters, the shape parameter (λ) and the mean parameter (μ), which collectively represent the irregularity of the samples. The NB distribution is a flexible distribution that provides a discrete analogue of a gamma distribution. The geometric distribution, itself the discrete analogue of the exponential distribution, is a special case of the NB distribution when $\lambda = 1$.

The parameters μ and λ represent different aspects of sampling irregularity, as illustrated by the examples shown in **Figure 2**. The mean parameter μ represents the fractional increase in the average interval between samples due to gaps: $\mu = \text{mean}(\Delta t^*) - 1 = (\Delta t_{average} - \Delta t_{nominal})/\Delta t_{nominal}$. Thus the special case of $\mu = 0$ corresponds to regular sampling (*i.e.*, $\Delta t_{average} = \Delta t_{nominal}$), whereas any larger value of μ corresponds to irregular sampling (*i.e.*, $\Delta t_{average} > \Delta t_{nominal}$) (**Figure 2c**). The shape parameter λ characterizes the similarity of gaps to each other; that is, a small λ indicates that the samples contain gaps of widely varying lengths, whereas a large λ indicates that the samples contain many gaps of similar lengths (**Figure 2a-2b**).

To visually illustrate these gap distributions, representative samples of irregular~~gappy~~ time series are presented in **Figure 1** for the three special processes described above (**Section 1.2**), *i.e.*, white noise, pink noise, and Brown noise. Specifically, three different gap distributions, namely, NB($\lambda = 1$, $\mu = 1$), NB($\lambda = 1$, $\mu = 14$), and NB($\lambda = 0.01$, $\mu = 1$), were simulated and each was applied to convert the three original (regular) time series (**Figure 1a-1c**) to irregular~~gappy~~ time series (**Figure 1d-1l**). These simulations clearly illustrate the effects of the two parameters λ and μ. In particular, compared with NB($\lambda = 1$, $\mu = 1$), NB($\lambda = 1$, $\mu = 14$) shows a similar level of sampling irregularity (same λ) but a much longer average gap interval (larger μ). Again compared with NB($\lambda = 1$, $\mu = 1$), NB($\lambda = 0.01$, $\mu = 1$) shows the same average interval (same μ)

but a much more irregular (skewed) gap distribution that contains a few very large gaps (smaller λ).

2.2. Examination of Sampling Irregularity in Real River Water-Quality Data

The above modeling approach was applied to real water-quality data from two large river monitoring networks in the United States to examine sampling irregularity. One such network is the Chesapeake Bay River Input Monitoring program, which typically samples streams bi-monthly to monthly, accompanied with additional sampling during stormflows (Langland *et al.*, 2012; Zhang *et al.*, 2015). These data were obtained from the U.S. Geological Survey National Water Information System (http://doi.org/10.5066/F7P55KJN). The other network is the Lake Erie and Ohio tributary monitoring program, which typically samples streams at a daily resolution (National Center for Water Quality Research, 2015). For each site, we determined the NB parameters to quantify sampling irregularity. The mean parameter μ can be estimated as described above, and the shape parameter λ can be calculated directly from the mean and variance of Δt^* as follows: $\lambda = \mu^2/[\mathrm{var}(\Delta t^*) - \mu] = (\mathrm{mean}(\Delta t^*) - 1)^2/[\mathrm{var}(\Delta t^*) - \mathrm{mean}(\Delta t^*) + 1]$. Alternatively, a maximum likelihood approach can be used, which employs the "*fitdist*" function in the "*fitdistrplus*" R package (Delignette-Muller and Dutang, 2015). In general, the two approaches have produced similar results, which are summarized in **Table 1**, with two examples of fitted NB distributions shown in **Figure 3**.

For the Chesapeake Bay River Input Monitoring program (9 sites), total nitrogen (TN) and total phosphorus (TP) are taken as representatives of water-quality constituents. According to the maximum likelihood approach, the shape parameter λ varies between 0.7 and 1.2 for TN and between 0.8 and 1.1 for TP (**Table 1**). These λ values are around 1.0, reflecting the fact that these sites have relatively even gap distributions (*i.e.*, relatively balanced counts of large and small gaps). The mean parameter μ varies between 9.5 and 19.6 for TN and between 13.4 and 24.4 for TP in the Chesapeake monitoring network, corresponding to $\Delta t_{average}$ of 10.5–20.6 days for TN and 14.4–25.4 days for TP, respectively. This is consistent with the fact that these sites have typically been sampled bi-monthly to monthly, along with additional sampling during stormflows (Langland *et al.*, 2012; Zhang *et al.*, 2015).

For the Lake Erie and Ohio tributary monitoring program (6 sites), the record of nitrate-plus-nitrite (NO_x) and TP were examined. According to the maximum likelihood approach, the shape parameter λ is approximately 0.01 for both constituents (**Table 1**). These very low λ values occur because these time series contain a few very large gaps, ranging from 35 days to 1109 days (~3 years). The mean parameter μ varies between 0.06 and 0.22, corresponding to $\Delta t_{average}$ of 1.06 and 1.22 days, respectively. This is consistent with fact that these sites have been sampled at a daily resolution with occasional missing values on some days (Zhang and Ball, 2017).

2.3. Simulation of Time Series with Irregular Sampling

To evaluate the various β estimation methods, our first step was to use Monte Carlo simulation to produce time series that mimic the sampling irregularity observed in real water-quality monitoring data. We began by simulating regular (gap free) time series using the fractional noise simulation method of Witt and Malamud (2013), which is based on inverse Fourier filtering of white noises. Our analysis showed this method performed reasonably well compared to other simulation methods for β values between 0 and 1 (~~data not shown~~see Supporting Information S1). In addition, this method can also simulate β values beyond this range. The noises simulated by the Witt and Malamud method, however, are band-limited to the Nyquist frequency (half of the sampling frequency) of the underlying white noise time series, whereas true fractional noises would contain spectral power at all frequencies, extending well above the Nyquist frequency for any sampling. Thus these band-limited noises will be less susceptible to spectral aliasing than true fractional noises would be; see Kirchner (2005) for detailed discussions of the aliasing issue.

~~Thirty~~100 replicates of regular (gap free) time series were produced for nine prescribed spectral slopes, which vary from $\beta = 0$ (white noise) to $\beta = 2$ (Brownian motion or "random walk") with an increment of 0.25 (*i.e.*, 0, 0.25, 0.5, 0.75, 1.0, 1.25, 1.5, 1.75, and 2). These regular time series each have a length (N) of 9125, which can be interpreted as 25 years of regular daily samples (that is, $\Delta t_{nominal} = 1$ day).

Each of the simulated regular time series was converted to irregular time series using gap intervals that were simulated with NB distributions. To make these gap intervals mimic those in typical river water-quality time series, representative NB parameters were chosen based on results from **Section 2.2**. Specifically, μ was set at 1 and 14, corresponding to $\Delta t_{average}$ of 2 days and 15 days respectively. For λ, we chose four values that span three orders of magnitude, *i.e.*, 0.001, 0.1, 1, and 10. Note that when $\lambda = 1$ the generated time series corresponds to a Bernoulli process. With the chosen values of μ and λ, a total of eight scenarios were generated, which were implemented using the "*rnbinom*" function in the "*stats*" R package (R Development Core Team, 2014):

1) $\mu = 1$ (*i.e.*, $\Delta t_{average}/\Delta t_{nominal} = 2$), $\lambda = 0.01$,
2) $\mu = 1$, $\lambda = 0.1$,
3) $\mu = 1$, $\lambda = 1$,
4) $\mu = 1$, $\lambda = 10$,
5) $\mu = 14$ (*i.e.*, $\Delta t_{average} / \Delta t_{nominal} = 15$), $\lambda = 0.01$,
6) $\mu = 14$, $\lambda = 0.1$,
7) $\mu = 14$, $\lambda = 1$,
8) $\mu = 14$, $\lambda = 10$.

Examples of these simulations are shown with boxplots in **Figure 2**.

3. Evaluation of Proposed Estimation Methods for Irregular Time Series

3.1. Summary of Estimation Methods

For the simulated irregular time series, β was estimated using the aforementioned two types of approaches. The first type includes 11 different interpolation methods (designated as B1-B11 below) to fill the data gaps, thus making the data regular and analyzable by traditional methods:

B1) Global mean: all missing values replaced with the mean of all observations;

B2) Global median: all missing values replaced with the median of all observations;

B3) Random replacement: all missing values replaced with observations randomly drawn (with replacement) from the time series;

B4) Next observation carried backward: each missing value replaced with the next available observation;

B5) Last observation carried forward: each missing value replaced with the preceding available observation;

B6) Average of the two nearest samples: it replaces each missing value with the mean of its next and preceding available observations;

B7) Lowess (locally weighted scatterplot smoothing) with a smoothing span of 1: missing values replaced using fitted values from a lowess model determined using all available observations (Cleveland, 1981);

B8) Lowess with a smoothing span of 0.75: same as B7 except that the smoothing span is 75% of the available data (similar distinction follows for B9-B11);

B9) Lowess with a smoothing span of 50%;

B10) Lowess with a smoothing span of 30%; and

B11) Lowess with a smoothing span of 10%.

B4 and B5 were implemented using the "*na.locf*" function in the "*zoo*" R package (Zeileis and Grothendieck, 2005). B7-B11 were implemented using the "*loess*" function in the "*stats*" R package (R Development Core Team, 2014). An illustration of these interpolation methods is provided in **Figure 4**. The interpolated data, along with the original regular data (designated as A1) were analyzed using the Whittle's maximum likelihood method for β estimation, which was implemented using the "*FDWhittle*" function in the "*fractal*" R package (Constantine and Percival, 2014).

The second type of approaches estimates β in the irregularly sampled data directly, using several variants of the Lomb-Scargle periodogram (designated as C1a-C1c below), and a recently developed wavelet-based method (designated as C2 below). Specifically, these approaches are:

C1a) Lomb-Scargle periodogram: the spectral density of the time series (with gaps) is estimated and the spectral slope is fit using all frequencies (Lomb, 1976; Scargle, 1982). This is a classic method for examining periodicity in irregularly sampled data, which is analogous to the more familiar fast Fourier transform method often used for regularly sampled data;

C1b) Lomb-Scargle periodogram with 5% data: same as C1a except that the fitting of the spectral slope considers only the lowest 5% frequencies (Montanari *et al.*, 1999);

C1c) Lomb-Scargle periodogram with "binned" data: same as C1a except that the fitting of the spectral slope is performed on binned data in three steps: (1) The entire range of frequency is divided into 100 equal-interval bins on logarithmic scale. (2) The respective medians of frequency and power spectral density are calculated for each of the 100 bins. (3) The 100 pairs of median frequency and median spectral density are used to estimate the spectral slope on a log-log scale.

C2) Kirchner and Neal (2013)'s wavelet method: uses a modified version of Foster's weighted wavelet spectrum (Foster, 1996) to suppress spectral leakage from low frequencies and applies an aliasing filter (Kirchner, 2005) to remove spectral aliasing artifacts at high frequencies.

C1a was implemented using the "*spec.ls*" function in the "*cts*" R package (Wang, 2013). C2 was run in *C*, using codes modified from those in Kirchner and Neal (2013).

3.2. Evaluation of Methods' Performance

Each estimation method listed above was applied to the simulated data (**Section 2.3**) to estimate β, which were then compared with the prescribed ("true") β to quantify the performance of each method. Plots of method evaluation for all simulations are provided as **Figures ~~S1~~S3-~~S10~~ S12** ~~in the~~ (Supporting Information S2). Close inspections of these plots reveal some general patterns of the methods' performance. For brevity, these patterns are presented with a subset of the plots, which correspond to the cases where true β = 1 and shape parameter λ = 0.01, 0.1, 1, and 10 (**Figure 5**). In general, β values estimated using the regular data (A1) are very close to 1.0, which indicates that the adopted fractional noise generation method and the Whittle's maximum likelihood estimator have small combined simulation and estimation bias. This is perhaps unsurprising, since the estimator is based on the Fourier transform and the noise generator is based on an inverse Fourier transform; thus, one method is essentially just the inverse of the other. One should also note that when fractional noises are not arbitrarily band-limited at the Nyquist frequency (as they inherently are with the noise generator that is used here), spectral aliasing should lead to spectral slopes that are flatter than expected (Kirchner, 2005), and thus to underestimates of LRD.

For the simulated irregular data, the estimation methods differ widely in their performance. Specifically, three interpolation methods (*i.e.*, B4-B6) consistently over-estimate β, indicating that they introduce additional correlations into the time series, reducing its short-timescale variability. In contrast, the other eight interpolation methods (*i.e.*, B1-B3 and B7-B11) generally under-estimate β, indicating that the interpolated points are less correlated than the original time series, thus introducing additional variability on short timescales. As expected, results from the lowess methods (B7-B11) depend strongly on the size of smoothing window, that is, more severe under-estimation of β is produced as the smoothing window becomes wider. In fact, when the smoothing window is 1.0 (*i.e.*, method B7), lowess performs the interpolation using all data available and thus behaves similarly to interpolations based on global means (B1) or global medians (B2), except that lowess fits a polynomial curve instead of constant values. However, whenever a sampling gap is much shorter than the smoothing window, the infilled lowess value will be close to the local mean or median, and the abrupt jumps produced by these infilled values will artificially increase the variance in the time series at high frequencies, leading to an artificially reduced spectral slope β and correspondingly, an underestimate of β. This mechanism explains why lowess interpolation distorts β more when there are many small gaps (large λ), and therefore more jumps to, and away from, the infilled values, than when there are only a few large gaps (small λ).

Among the direct methods (*i.e.*, C1a, C1b, C1c, and C2), the Lomb-Scargle method, with original data (C1a) or binned data (C1c) tends to under-estimate β, though the underestimation by C1c is generally less severe. The modified Lomb-Scargle method (C1b), using only the lowest 5% of frequencies, yields estimates that are centered around 1.0. However, C1b has the highest variability (*i.e.*, least precision) in β estimates among all methods. Compared with all the above methods, the wavelet method (C2) has much better performance in terms of both accuracy and precision when λ is 1 or 10, a slightly better performance when λ is 0.1, but a worse performance when λ is 0.01.

The shape parameter λ greatly affects the performance of the estimation methods. All the interpolation methods that under-estimate β (*i.e.*, B1-B3 and B7-B11) perform worse as λ increases from 0.01 to 10. This effect can be interpreted as follows: when the time series contains a large number of relatively small gaps (*e.g.*, λ = 1 or 10), there are many jumps (which, as noted above, contain mostly high-frequency variance) between the original data and the infilled values, resulting in more severe under-estimation. In contrast, when the data contain only a small number of very large gaps (*e.g.*, λ = 0.01 or 0.1), there are fewer of these jumps, resulting in minimal under-estimation. Similar effects of λ are also observed with the interpolation methods that show over-estimation (*i.e.*, B4-B6) – that is, over-estimation is more severe when λ is larger. Similarly, the Lomb-Scargle method (C1a and C1c) performs worse (more serious underestimation) as λ increases. Finally, method C2 seems to perform the best when λ is large (1 or 10), but not well when λ is very small (0.01), as noted above. This result highlights the sensitivity of the wavelet method to the presence of a few large gaps in the time series. For such cases, a potentially more feasible approach is to break the whole time series into several segments (each without long gaps) and then apply the wavelet method (C2) to analyze each segment separately. If this can yield more accurate estimates, then further simulation experiments should be designed to systematically determine how long the gap needs to be to invoke such an approach.

Next, the method evaluation is extended to all the simulated spectral slopes, that is, β = 0, 0.25, 0.5, 0.75, 1.0, 1.25, 1.5, 1.75, and 2. For ease of discussion, three quantitative criteria were proposed for evaluating performance, namely, bias (B), standard deviation (SD), and root-mean-squared error (RMSE), as defined below:

$$B_i = \overline{\beta_i} - \beta_{true} \tag{14}$$

$$SD_i = \sqrt{\frac{1}{99\cancel{29}} \sum_{j=1}^{100\cancel{30}} (\beta_{i,j} - \overline{\beta_i})^2} \tag{15}$$

$$RMSE_i = \sqrt{B_i^2 + SD_i^2} \tag{16}$$

where $\overline{\beta_i}$ is the mean of ~~30~~100 β values estimated by method i, and β_{true} is the prescribed β value for simulation of the initial regular time series. In general, B and SD can be considered as the models' systematic error and random error, respectively, and RMSE serves as an integrated measure of both errors. For all evaluations, plots of bias and RMSE are provided in the main text. (Plots of SD are provided as **Figure ~~S5~~ S7** and **Figure ~~S10~~ S12** for simulations with μ = 1 and μ = 14, respectively.)

For simulations with μ = 1, results of estimation bias and RMSE are summarized in **Figure 6** and **Figure 7**, respectively. (More details are provided in **Figures ~~S1~~S3-~~S4~~S6**.) For brevity, we focus on three direct methods (C1a, C1b and C2) and three representative interpolation methods. (Specifically, B1 represents B1-B3 and B7; B6 represents B4-B6, and B8 represents B8-B11.) Overall, these six methods show mixed performances. In terms of bias (**Figure 6**), B1 (global mean) and B8 (lowess with a smoothing span of 0.75) tend to have negative bias, particularly for time series with (1) moderate-to-large β_{true} values and (2) large λ values (*i.e.*, less skewed gap intervals). By contrast, B1 and B8 generally have minimal bias when (1) β_{true} is close to zero (*i.e.*, when the simulated time series is close to white noise); and (2) λ is small (*e.g.*, 0.01), since interpolating a few large gaps cannot significantly affect the overall correlation structure. In addition, lowess interpolation with a larger smoothing window tends to yield more negatively biased estimates (data not shown). The other interpolation method, B6 (mean of the two nearest neighbors) tends to over-estimate β, particularly for time series with (1) small β_{true} values and (2) large λ values. At large β_{true} values (*e.g.*, 2.0), the auto-correlation is already very strong such that taking the mean of two neighbors for gap filling does not introduce much additional correlation, as opposed to the case of small β_{true} values. The Lomb-Scargle methods (C1a and C1b) generally have negative bias, particularly for time series with (1) moderate-to-large β_{true} values (for both methods) and (2) large λ values (for C1a), which is similar to B1 and B8. However, C1b overall shows less severe bias than C1a. Finally, the wavelet method (C2) shows generally the smallest bias among all methods. However, its performance advantage is not as great when the time series has small λ values (*i.e.*, very skewed gap intervals), as noted above, which may be due to the fact that the aliasing filter was designed for regular time series. In terms of SD (**Figure ~~S5~~S7**), method C1b performs the worst among all methods (as noted above), method B6 and B8 perform poorly for large β_{true} values, and method C2 performs poorly for β_{true} = 0. In terms of RMSE (**Figure 7**), methods B1, B8, C1a, and C1b perform well for small β_{true} values and small λ values, whereas method B6 performs well for large β_{true} values and small λ values. In comparison, method C2 has the smallest RMSEs among all methods, and its RMSEs are similarly small for the wide range of β_{true} and λ values. In general, the wavelet method can be considered the best among all methods.

For simulations with μ = 14, results of estimation bias and RMSE are summarized in **Figure 8** and **Figure 9**, respectively. (More details are provided in **Figures ~~S6~~S8-~~S9~~S11**.) Overall, these methods show mixed performances that are generally similar to the cases when μ = 1, as discussed above. These results highlight the generality of these methods' performances, which applies at least to the range of μ = [1, 14]. In addition, all methods show generally larger RMSE for μ = 14 than μ = 1, indicating their dependence on the mean gap interval (**Figure 9**). Perhaps the most notable difference is observed with method C2, which in this case shows positive bias for small λ values (0.01 and 0.1) and negative bias for large λ values (1 and 10) (**Figure 8f**). It nonetheless generally shows the smallest RMSEs among all the tested methods.

3.3. Quantification of Spectral Slopes in Real Water-Quality Data

In this section, the proposed estimation approaches were applied to quantify β in real water-quality data from the two monitoring programs presented in **Section 2.2** (**Table 1**). As noted in **Section 1.3**, such real data are typically much more complex than our simulated time series, because of (1) strong deviations from normal distributions and (2) effects of flow-dependence, seasonality, and temporal trend (Hirsch *et al.*, 1991; Helsel and Hirsch, 2002). In this regard, future research may simulate time series with these important characteristics and evaluate the performance of various estimation approaches, perhaps following the modeling framework described herein. Alternatively, one may quantify β in transformed time series after accounting for the above aspects. In this work, we have taken the latter approach for a preliminary investigation. Specifically, we have used the published Weighted Regressions on Time, Discharge, and Season (WRTDS) method (Hirsch *et al.*, 2010) to transform the original time series. This widely accepted method estimates daily concentrations based on discretely collected concentration samples using time, season, and discharge as explanatory variables, *i.e.*,

$$ln(C) = \beta_0 + \beta_1 t + \beta_2 ln(Q) + \beta_3 sint(2\pi t) + \beta_4 cos(2\pi t) + \varepsilon \quad (17)$$

where *C* is concentration, *Q* is daily discharge, *t* is time in decimal years, β_i are fitted coefficients, and ε is the error term. The 2nd and 3rd terms on the right represent time and discharge effects, respectively, whereas the 4th and 5th terms collectively represent cyclical seasonal effects. For a full description of this method, see Hirsch *et al.* (2010) ~~Hirsch and De Cicco (2015)~~. In this work, WRTDS was applied to obtain the time series of estimated daily concentration for each constituent at each site. The difference between observed concentration (C_{obs}) and estimated concentration (C_{est}) was calculated in logarithmic space to obtain the concentration residuals,

$$residuals = ln(C_{obs}) - ln(C_{est}) \quad (18)$$

For our data sets, histograms of concentration residuals (expressed in natural log concentration units) are shown in **Figures ~~S11~~S13-~~S14~~S16** (Supporting Information S3). Compared with the original concentration data, these model residuals are much more nearly normal and homoscedastic. Moreover, the model residuals are less susceptible to the issues of temporal, seasonal, and discharge-drive variations than the original concentrations. Therefore, the model residuals are more appropriate than the original concentrations for β estimation using the simulation framework adopted in this work.

The estimated β values for the concentration residuals are summarized in **Figure 10**. Clearly, the estimated β varies considerably with the estimation method. In addition, the estimated β varies with site and constituent (*i.e.*, TP, TN, or NO_x.) Our discussion below focuses on the wavelet method (C2), because it is established above that this method performs better than the other estimation methods under a wide range of gap conditions. We emphasize that it is beyond our current scope to precisely quantify β in these water-quality data sets, but our simulation results presented above (**Section 3.2**) can be used as references to qualitatively evaluate the reliability of C2 and/or other methods for these data sets.

For TN and TP concentration data at the Chesapeake River Input Monitoring sites (**Table 1**), μ varies between 9.5 and 24.4, whereas λ is ~1.0. Thus, the simulated gap scenario of NB(μ = 14, λ = 1) can be used as a reasonable reference to assess methods' reliability (**Figure 8**). Based on method C2, the estimated β ranges between β = 0.36 and β = 0.61 for TN and between β = 0.30 and β = 0.58 for TP at these sites (**Figure 10**). For such ranges, the simulation results indicate that method C2 tends to moderately under-estimate β under this gap scenario (**Figure 8**), and hence spectral slopes for TN and TP at these Chesapeake sites are likely slightly higher than those presented above.

For NO_x and TP concentration data at the Lake Erie and Ohio sites (**Table 1**), μ varies between 0.06 and 0.22, whereas λ is ~0.01. Thus, the simulated gap scenario of NB(μ = 1, λ = 0.01) can be used as a reasonable reference to assess the methods' reliability (**Figure 6**). For such small λ (*i.e.*, a few gaps that are very dissimilar from others), C2 is not reliable for β estimation, as reflected by the generally positive bias in the simulation results. By contrast, methods B1 (interpolation with global mean) and B8 (lowess with span 0.75) both perform quite well under this gap scenario (**Figure 6**). These two methods provide almost identical β estimates for each site-constituent combination, ranging from β = 1.0 to β = 1.5 for NO_x and from β = 1.0 to β = 1.4 for TP (**Figure 10**).

Overall, the above analysis of real water-quality data has illustrated the wide variability in β estimates, with different choices of estimation methods yielding very different results. To our knowledge, these water-quality data have not heretofore been analyzed in this context. As illustrated above, our simulation experiments (**Section 3.2**) can be used as references to coarsely evaluate the reliability of each method under specific gap scenarios, thereby considerably narrowing the likely range of the estimated spectral slopes. Nonetheless, our results demonstrate that the analyzed water-quality time series can exhibit strong fractal scaling, particularly at the Lake Erie and Ohio tributary sites. Thus, an important implication is that researchers and analysts should be cautious when applying standard statistical methods to identify temporal trends in such water-quality data sets (Kirchner and Neal, 2013). In future work, one may consider applying Bayesian statistical analysis or other approaches to more accurately quantify the spectral slope and associated uncertainty for real water-quality data analysis. In addition, the modeling framework presented herein (including both gap simulation and β estimation) may be extended to simulations of irregular time series that have prescribed spectral slopes and also superimposed temporal trends, which can then be used to evaluate the validity of various statistical methods for identifying trend and associated statistical significance.

4. Conclusions

River water-quality time series often exhibit fractal scaling behavior, which presents challenges to the identification of deterministic trends. Because traditional estimation methods are generally not applicable to irregularly sampled time series, we have examined two broad types of estimation approaches and evaluated their performances against synthetic data with a wide range of prescribed β values and gap intervals representative of the sampling irregularity of real water-quality data.

The results of this work suggest several important messages. First, the results remind us of the risks in using interpolation for gap filling when examining auto-correlation, as the interpolation methods consistently under-estimate or over-estimate β under a wide range of prescribed β values and gap distributions. Second, the long-established Lomb-Scargle spectral method also consistently under-estimates β. Its modified form, using the 5% lowest frequencies for spectral slope estimation, has very poor precision, although the overall bias is small. Third, the wavelet method, coupled with an aliasing filter, has the smallest bias and root-mean-squared error among all methods for a wide range of prescribed β values and gap distributions, except for cases with small prescribed β values (*i.e.*, close to white noise) or small λ values (*i.e.*, very skewed gap distributions). Thus, the wavelet method is recommended for estimating spectral slope in irregular time series until improved methods are developed. In this regard, future research should aim to develop an aliasing filter that is more applicable to irregular time series with very skewed gap intervals. Finally, all methods' performances depend strongly on the sampling irregularity in terms of both the skewness and mean of gap-interval lengths, highlighting that the accuracy and precision of each method are data-specific.

Overall, these results provide new contributions in terms of better understanding and quantification of the proposed methods' performances for estimating the strength of fractal scaling in irregularly sampled water-quality data. In addition, the work has provided an innovative and general approach for modeling sampling irregularity in water-quality records.

Moreover, this work has proposed and demonstrated a generalizable framework for data simulation (with gaps) and β estimation, which can be readily applied toward the evaluation of other methods that are not covered in this work. More generally, the findings and approaches may also be broadly applicable to irregularly sampled data in other scientific disciplines. Last but not least, we note that accurate quantification of fractal scaling in irregular water-quality time series remains an unresolved challenge for the hydrologic community and for many other disciplines that must grapple with irregular sampling.

Data Availability

River monitoring data used in this study are available through the U.S. Geological Survey National Water Information System (http://doi.org/10.5066/F7P55KJN) and the Heidelberg University's National Center for Water Quality Research.

Supporting Information

Supporting information to this article is available online.

Competing Interests

The authors declare that they have no conflict of interest.

Acknowledgements

Zhang was supported by the Maryland Sea Grant through awards NA10OAR4170072 and NA14OAR1470090 and by the Maryland Water Resources Research Center through a graduate fellowship while he was a doctoral student at the Johns Hopkins University. Subsequent support to Zhang was provided by the USEPA under grant "EPA/CBP Technical Support 2017" (No. 07-5-230480). Harman's contribution to this work was supported by the National Science Foundation through grants CBET-1360415 and EAR-1344664. We thank Bill Ball (Johns Hopkins University) and Bob Hirsch (U.S. Geological Survey) for many useful discussions. This is contribution no. xxxx to the University of Maryland Center for Environmental Science.

References

Aubert, A. H., J. W. Kirchner, C. Gascuel-Odoux, M. Faucheux, G. Gruau and P. Mérot, 2014. Fractal water quality fluctuations spanning the periodic table in an intensively farmed watershed. *Environ. Sci. Technol.* 48:930-937, DOI: 10.1021/es403723r.

Beran, J., 2010. Long-range dependence. *Wiley Interdiscip. Rev. Comput. Stat.*~~*Wiley Interdisciplinary Reviews: Computational Statistics*~~ 2:26-35, DOI: 10.1002/wics.52.

Beran, J., Y. Feng, S. Ghosh and R. Kulik, 2013. *Long-Memory Processes: Probabilistic Properties and Statistical Methods*. Berlin, Heidelberg, Springer Berlin Heidelberg, ISBN 978-3-642-35511-0

Boutahar, M., V. Marimoutou and L. Nouira, 2007. Estimation Methods of the Long Memory Parameter: Monte Carlo Analysis and Application. *J. Appl. Stat.*~~*Journal of Applied Statistics*~~ 34:261-301, DOI: 10.1080/02664760601004874.

Box, G. E. P., G. M. Jenkins and G. C. Reinsel, 2008. *Time Series Analysis, Fourth Edition*. Hoboken, NJ, John Wiley & Sons, Inc., ISBN 9781118619193

Clarke, R. T., 2013. Calculating uncertainty in regional estimates of trend in streamflow with both serial and spatial correlations. *Water Resour. Res.* 49:7120-7125, DOI: 10.1002/wrcr.20465.

Cleveland, W. S., 1981. LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *Am. Stat.* ~~*American Statistician*~~ 35:54, DOI: 10.2307/2683591.

Cohn, T. A. and H. F. Lins, 2005. Nature's style: Naturally trendy. *Geophys. Res. Lett.* 32:L23402, DOI: 10.1029/2005GL024476.

Constantine, W. and D. Percival, 2014. *fractal: Fractal Time Series Modeling and Analysis*,

Darken, P. F., C. E. Zipper, G. I. Holtzman and E. P. Smith, 2002. Serial correlation in water quality variables: Estimation and implications for trend analysis. *Water Resour. Res.* 38:1117, DOI: 10.1029/2001WR001065.

Delignette-Muller, M. L. and C. Dutang, 2015. fitdistrplus: An R Package for Fitting Distributions. *J. Stat. Softw.*~~*Journal of Statistical Software*~~ 64:1-34, DOI.

Ehsanzadeh, E. and K. Adamowski, 2010. Trends in timing of low stream flows in Canada: impact of autocorrelation and long-term persistence. *Hydrol. Process.* 24:970-980, DOI: 10.1002/hyp.7533.

Fatichi, S., S. M. Barbosa, E. Caporali and M. E. Silva, 2009. Deterministic versus stochastic trends: Detection and challenges. *J. Geophys. Res.*~~*Journal of Geophysical Research*~~ 114:D18121, DOI: 10.1029/2009JD011960.

Foster, G., 1996. Wavelets for period analysis of unevenly sampled time series. ~~*The Astronomical Journal*~~ *Astron. J.* 112:1709-1729, DOI, http://articles.adsabs.harvard.edu/full/1996AJ....112.1709F.

Franzke, C., 2012a. Nonlinear Trends, Long-Range Dependence, and Climate Noise Properties of Surface Temperature. *J. Clim.* 25:4172-4183, DOI: 10.1175/JCLI-D-11-00293.1.

Franzke, C., 2012b. On the statistical significance of surface air temperature trends in the Eurasian Arctic region. *Geophys. Res. Lett.* 39:L23705, DOI: 10.1029/2012GL054244.

Godsey, S. E., W. Aas, T. A. Clair, H. A. de Wit, I. J. Fernandez, J. S. Kahl, I. A. Malcolm, C. Neal, M. Neal, S. J. Nelson, S. A. Norton, M. C. Palucis, B. L. Skjelkvåle, C. Soulsby, D. Tetzlaff and J. W. Kirchner, 2010. Generality of fractal 1/f scaling in catchment tracer time series, and its implications for catchment travel time distributions. *Hydrol. Process*. 24:1660-1671, DOI: 10.1002/hyp.7677.

Graham, J., 2009. Missing Data Analysis: Making It Work in the Real World. *Annu. Rev. Psychol.* 60:549-576, DOI: 10.1146/annurev.psych.58.110405.085530.

Helsel, D. R. and R. M. Hirsch, 2002. Statistical Methods in Water Resources. *U.S. Geological Survey Techniques of Water-Resources Investigations Book 4, Chapter A3*. U.S. Geological Survey, Reston, VA, p. 522. http://pubs.usgs.gov/twri/twri4a3/.

Hirsch, R. M., R. B. Alexander and R. A. Smith, 1991. Selection of methods for the detection and estimation of trends in water quality. *Water Resour. Res.* 27:803-813, DOI: 10.1029/91WR00259.

~~Hirsch, R. M. and L. De Cicco, 2015. User guide to Exploration and Graphics for RivEr Trends (EGRET) and dataRetrieval: R packages for hydrologic data (version 2.0, February 2015). U.S. Geological Survey Techniques and Methods Book 4, Chapter A10, Reston, VA, p. 93. http://dx.doi.org/10.3133/tm4A10.~~

Hirsch, R.M., D.L. Moyer and S.A. Archfield, 2010. Weighted regressions on time, discharge, and season (WRTDS), with an application to Chesapeake Bay river inputs. *J. Am. Water Resour. Assoc.* 46:857-880, DOI: 10.1111/j.1752-1688.2010.00482.x.

Hurst, H. E., 1951. Long-term storage capacity of reservoirs. *Trans. Amer. Soc. Civil Eng.* 116:770-808, DOI.

Khaliq, M. N., T. B. M. J. Ouarda and P. Gachon, 2009. Identification of temporal trends in annual and seasonal low flows occurring in Canadian rivers: The effect of short- and long-term persistence. ~~*Journal of Hydrology*~~*J. Hydro.* 369:183-197, DOI: 10.1016/j.jhydrol.2009.02.045.

Khaliq, M. N., T. B. M. J. Ouarda, P. Gachon and L. Sushama, 2008. Temporal evolution of low-flow regimes in Canadian rivers. *Water Resour. Res.* 44:W08436, DOI: 10.1029/2007WR006132.

Kirchner, J., 2005. Aliasing in 1⁄f^{α} noise spectra: Origins, consequences, and remedies. *Phys. Rev. E*~~*Physical Review E*~~ 71:066110-066110, DOI: 10.1103/PhysRevE.71.066110.

Kirchner, J. W., X. Feng and C. Neal, 2000. Fractal stream chemistry and its implications for contaminant transport in catchments. *Nature* 403:524-527, DOI: 10.1038/35000537.

Kirchner, J. W., X. Feng and C. Neal, 2001. Catchment-scale advection and dispersion as a mechanism for fractal scaling in stream tracer concentrations. ~~*Journal of Hydrology*~~*J. Hydro.* 254:82-101, DOI: 10.1016/s0022-1694(01)00487-5.

Kirchner, J. W. and C. Neal, 2013. Universal fractal scaling in stream chemistry and its implications for solute transport and water quality trend detection. *Proc. Natl. Acad. Sci. U. S. A.* 110:12213-12218, DOI: 10.1073/pnas.1304328110.

Kirchner, J. W. and A. Weil, 1998. No fractals in fossil extinction statistics. *Nature* 395:337-338, DOI: 10.1038/26384.

Langland, M. J., J. D. Blomquist, D. L. Moyer and K. E. Hyer, 2012. Nutrient and suspended-sediment trends, loads, and yields and development of an indicator of streamwater quality at nontidal sites in the Chesapeake Bay watershed, 1985-2010. U.S. Geological Survey Scientific Investigations Report 2012-5093, Reston, VA, p. 26. http://pubs.usgs.gov/sir/2012/5093/pdf/sir2012-5093.pdf.

Lennartz, S. and A. Bunde, 2009. Trend evaluation in records with long-term memory: Application to global warming. *Geophys. Res. Lett.* 36:L16706, DOI: 10.1029/2009GL039516.

Lomb, N. R., 1976. Least-squares frequency analysis of unequally spaced data. *Astrophysics and Space Science* 39:447-462, DOI: 10.1007/BF00648343.

Montanari, A., R. Rosso and M. S. Taqqu, 1997. Fractionally differenced ARIMA models applied to hydrologic time series: Identification, estimation, and simulation. *Water Resour. Res.* 33:1035-1044, DOI: 10.1029/97WR00043.

Montanari, A., R. Rosso and M. S. Taqqu, 2000. A seasonal fractional ARIMA Model applied to the Nile River monthly flows at Aswan. *Water Resour. Res.* 36:1249-1259, DOI: 10.1029/2000WR900012.

Montanari, A., M. S. Taqqu and V. Teverovsky, 1999. Estimating long-range dependence in the presence of periodicity: An empirical study. *Math. Comput. Model.*~~*Mathematical and Computer Modelling*~~ 29:217-228, DOI: 10.1016/S0895-7177(99)00104-1.

National Center for Water Quality Research, 2015. Tributary Data Download. Accessed July 23, 2015, https://www.heidelberg.edu/academics/research-and-centers/national-center-for-water-quality-research/tributary-data-download.

Noguchi, K., Y. R. Gel and C. R. Duguay, 2011. Bootstrap-based tests for trends in hydrological time series, with application to ice phenology data. ~~*Journal of Hydrology*~~*J. Hydro.* 410:150-161, DOI: 10.1016/j.jhydrol.2011.09.008.

R Development Core Team, 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3900051070. http://www.r-project.org.

Rea, W., L. Oxley, M. Reale and J. Brown, 2009. Estimators for Long Range Dependence: An Empirical Study. *Electron. J. Stat.*~~*Electronic Journal of Statistics*~~, http://arxiv.org/abs/0901.0762, DOI.

Sang, Y.-F., Z. Wang and C. Liu, 2014. Comparison of the MK test and EMD method for trend identification in hydrological time series. ~~*Journal of Hydrology*~~*J. Hydro.* 510:293-298, DOI: 10.1016/j.jhydrol.2013.12.039.

Scargle, J. D., 1982. Studies in Astronomical Time-Series Analysis. II. Statistical Aspects of Spectral-Analysis of Unevenly Spaced Data. *Astrophys. J.*~~*Astrophysical Journal*~~ 263:835-853, DOI: 10.1086/160554.

Stroe-Kunold, E., T. Stadnytska, J. Werner and S. Braun, 2009. Estimating long-range dependence in time series: an evaluation of estimators implemented in R. *Behav. Res. Methods* 41:909-923, DOI: 10.3758/BRM.41.3.909.

Taqqu, M. S., V. Teverovsky and W. Willinger, 1995. Estimators for long-range dependence: an empirical study. *Fractals*, http://www.worldscientific.com/doi/abs/10.1142/S0218348X95000692, DOI.

Wang, Z., 2013. cts: An R Package for Continuous Time Autoregressive Models via Kalman Filter. ~~*Journal of Statistical Software*~~*J. Stat. Softw.* 53:1-19, DOI.

Witt, A. and B. D. Malamud, 2013. Quantification of Long-Range Persistence in Geophysical Time Series: Conventional and Benchmark-Based Improvement Techniques. *Surv. Geophys.* 34:541-651, DOI: 10.1007/s10712-012-9217-8.

Yue, S., P. Pilon, B. Phinney and G. Cavadias, 2002. The influence of autocorrelation on the ability to detect trend in hydrological series. *Hydrol. Process.* 16:1807-1829, DOI: 10.1002/hyp.1095.

Zeileis, A. and G. Grothendieck, 2005. zoo: S3 Infrastructure for Regular and Irregular Time Series. *J. Stat. Softw.*~~*Journal of Statistical Software*~~ 14:1-27, DOI.

Zetterqvist, L., 1991. Statistical Estimation and Interpretation of Trends in Water Quality Time Series. *Water Resour. Res.* 27:1637-1648, DOI: 10.1029/91wr00478.

Zhang, Q. and W. P. Ball, 2017. Improving Riverine Constituent Concentration and Flux Estimation by Accounting for Antecedent Discharge Conditions. ~~*Journal of Hydrology*~~*J. Hydro.* 547:387–402, DOI: 10.1016/j.jhydrol.2016.12.052.

Zhang, Q., D. C. Brady, W. R. Boynton and W. P. Ball, 2015. Long-Term Trends of Nutrients and Sediment from the Nontidal Chesapeake Watershed: An Assessment of Progress by River and Season. *J. Am. Water Resour. Assoc.*~~*JAWRA Journal of the American Water Resources Association*~~ 51:1534-1555, DOI: 10.1111/1752-1688.12327.

Table 1. Quantification of sampling irregularity for selected water-quality constituents at nine sites of the Chesapeake Bay River Input Monitoring program and six sites of the Lake Erie and Ohio tributary monitoring program. (μ: mean parameter; λ: shape parameter estimated using maximum likelihood; λ': shape parameter estimated using the direct approach (see **Section 2.2**). $\Delta t_{average}$: average gap interval; N: total number of samples.)

I. Chesapeake Bay River Input Monitoring program

Site ID	River and station name	Drainage area (mi^2)	Total nitrogen (TN)					Total phosphorus (TP)				
			λ	λ'	μ	$\Delta t_{average}$ (days)	N	λ	λ'	μ	$\Delta t_{average}$ (days)	N
01578310	Susquehanna River at Conowingo, MD	27,100	0.8	1.1	13.5	14.5	876	0.8	1.0	13.4	14.4	881
01646580	Potomac River at Chain Bridge, Washington D.C.	11,600	0.9	0.6	9.5	10.5	1,385	1.1	1.0	24.4	25.4	579
02035000	James River at Cartersville, VA	6,260	0.8	1.0	13.9	14.9	960	0.8	1.1	13.7	14.7	974
01668000	Rappahannock River near Fredericksburg, VA	1,600	0.8	0.6	15.6	16.6	776	0.8	0.6	15.2	16.2	796
02041650	Appomattox River at Matoaca, VA	1,340	0.8	0.8	15.1	16.1	798	0.8	0.8	14.9	15.9	810
01673000	Pamunkey River near Hanover, VA	1,071	0.8	0.9	15.1	16.1	873	0.8	1.0	14.7	15.7	894
01674500	Mattaponi River near Beulahville, VA	601	0.7	0.9	14.3	15.3	810	0.8	0.9	14.2	15.2	820
01594440	Patuxent River at Bowie, MD	348	0.9	1.1	15.3	16.3	787	0.8	0.8	14.0	15.0	861
01491000	Choptank River near Greensboro, MD	113	1.2	1.5	19.6	20.6	680	1.1	1.0	20.5	21.5	690

II. Lake Erie and Ohio tributary monitoring program

Site ID	River and station name	Drainage area (mi^2)	Nitrate-plus-nitrite (NO_x)					Total phosphorus (TP)				
			λ	λ'	μ	$\Delta t_{average}$ (days)	N	λ	λ'	μ	$\Delta t_{average}$ (days)	N
04193500	Maumee River at Waterville, OH	6,330	0.005	0.0003	0.19	1.19	9,101	0.005	0.0003	0.19	1.19	9,101
04198000	Sandusky River near Fremont, OH	1,253	0.01	0.003	0.22	1.22	9,641	0.01	0.003	0.22	1.22	9,655
04208000	Cuyahoga River at Independence, OH	708	0.007	0.006	0.13	1.13	7,421	0.007	0.006	0.13	1.13	7,426
04212100	Grand River near Painesville, OH	686	0.01	0.005	0.21	1.21	5,023	0.01	0.005	0.22	1.22	4,994
04197100	Honey Creek at Melmore, OH	149	0.007	0.005	0.06	1.06	9,914	0.007	0.005	0.06	1.06	9,914
04197170	Rock Creek at Tiffin, OH	34.6	0.007	0.008	0.06	1.06	8,422	0.007	0.008	0.06	1.06	8,440

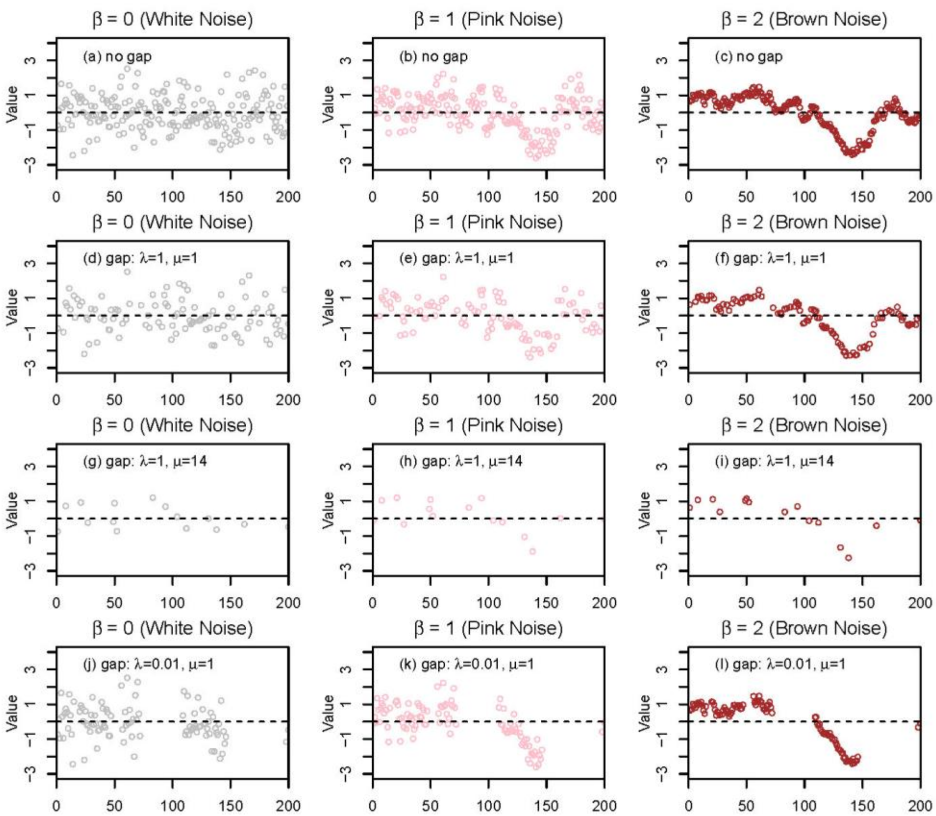

Figure 1. Synthetic time series with 200 time steps for three representative fractal scaling processes that correspond to white noise ($\beta = 0$), pink noise ($\beta = 1$), and Brown noise ($\beta = 2$). The 1st row shows the simulated time series without any gap. The three rows below show the same time series as in the 1st row but with data gaps that were simulated using three different negative binomial (NB) distributions, that is, 2nd row: NB($\lambda = 1$, $\mu = 1$); 3rd row: NB($\lambda = 1$, $\mu = 14$); 4th row: NB($\lambda = 0.01$, $\mu = 1$).

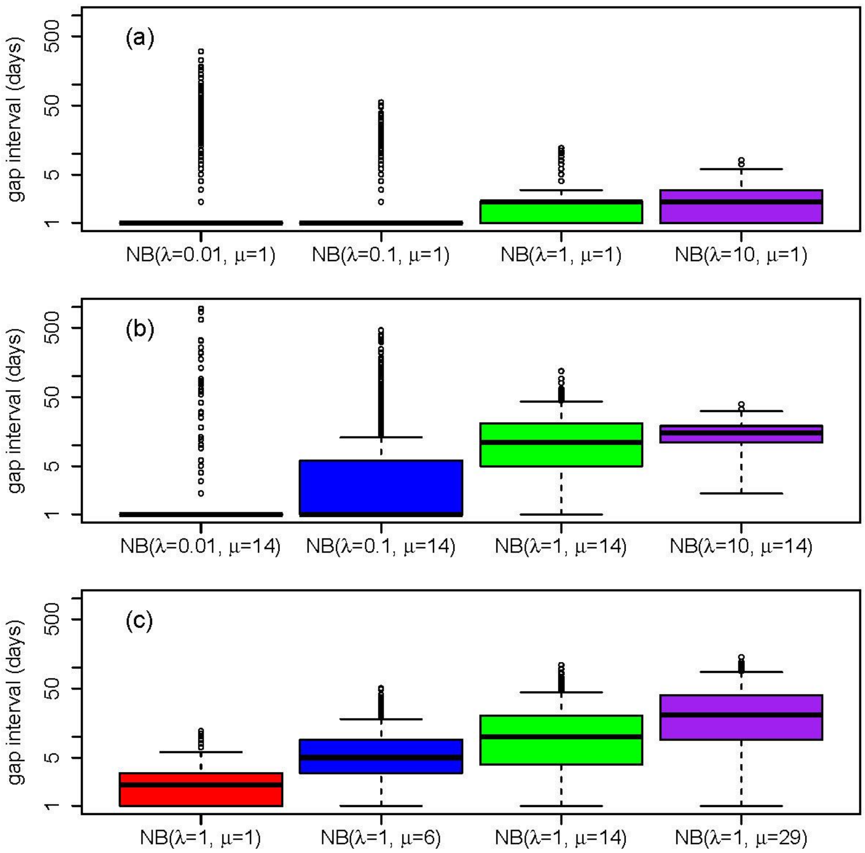

Figure 2. Examples of gap interval simulation using binomial distributions, NB (shape λ, mean μ). Simulation parameters: L = 9125 days, $\Delta t_{nominal}$ = 1 day. The three panels show simulation with fixed (a) μ = 1, (b) μ = 14, and (c) λ = 1. Note that $\Delta t_{average}/\Delta t_{nominal} = \mu + 1$.

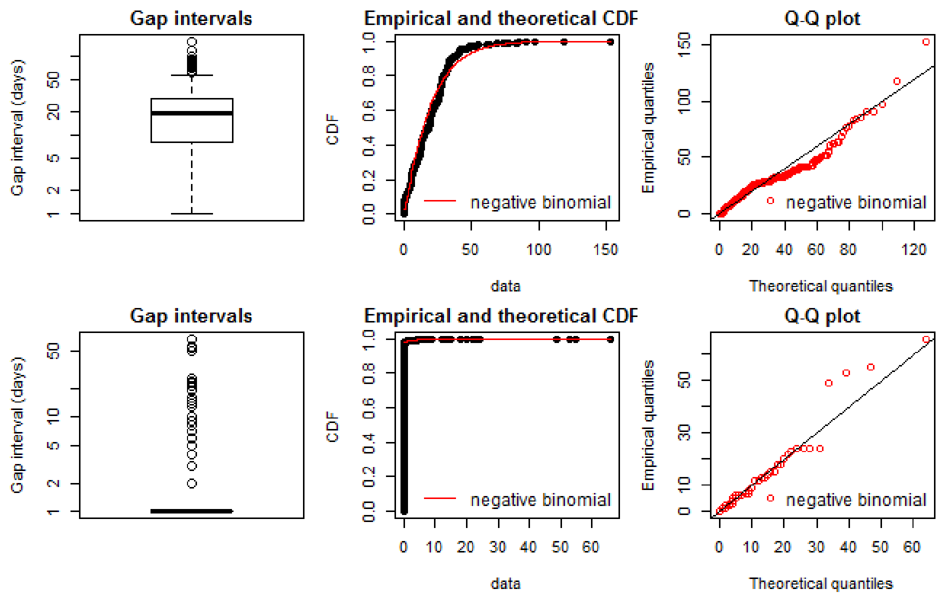

Figure 3. Examples of quantified sampling irregularity with negative binomial (NB) distributions: total nitrogen in Choptank River (top) and total phosphorus in Cuyahoga River (bottom). Theoretical CDF and quantiles are based on the fitted NB distributions. See **Table 1** for estimated mean and shape parameters.

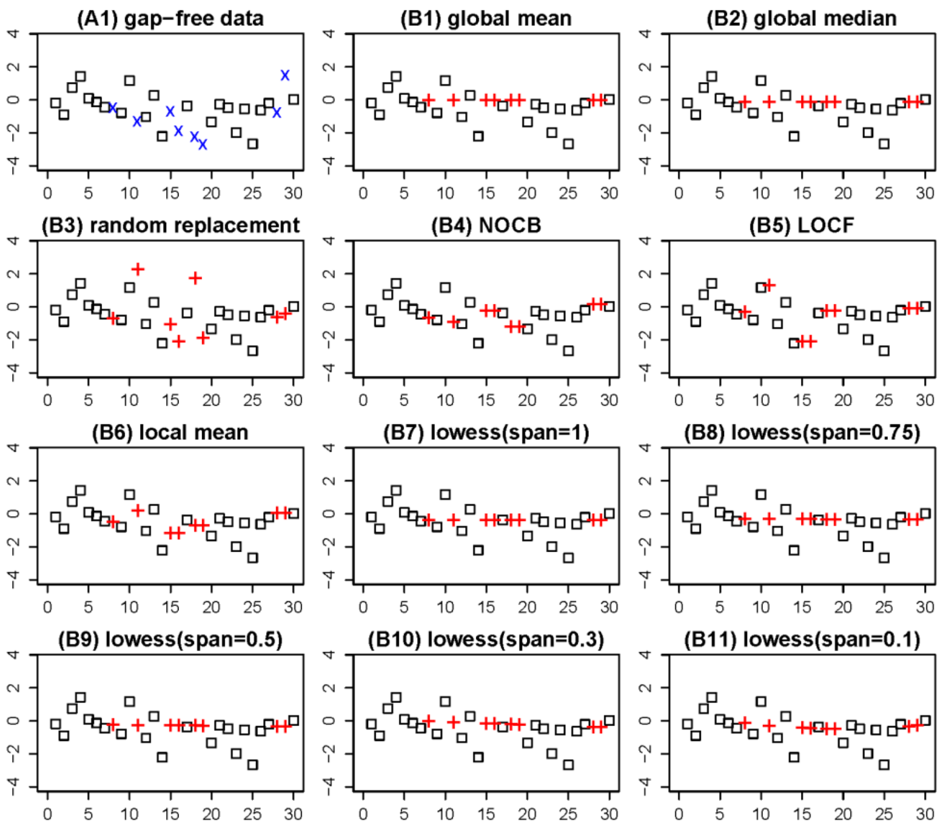

Figure 4. Illustration of the interpolation methods for gap filling. The gap-free data (A1) was simulated with a series length of 500, with the first 30 data shown. (x: omitted data for gap filling; +: interpolated data; NOCB: next observation carried backward; LOCF: last observation carried forward; lowess: locally weighted scatterplot smoothing.)

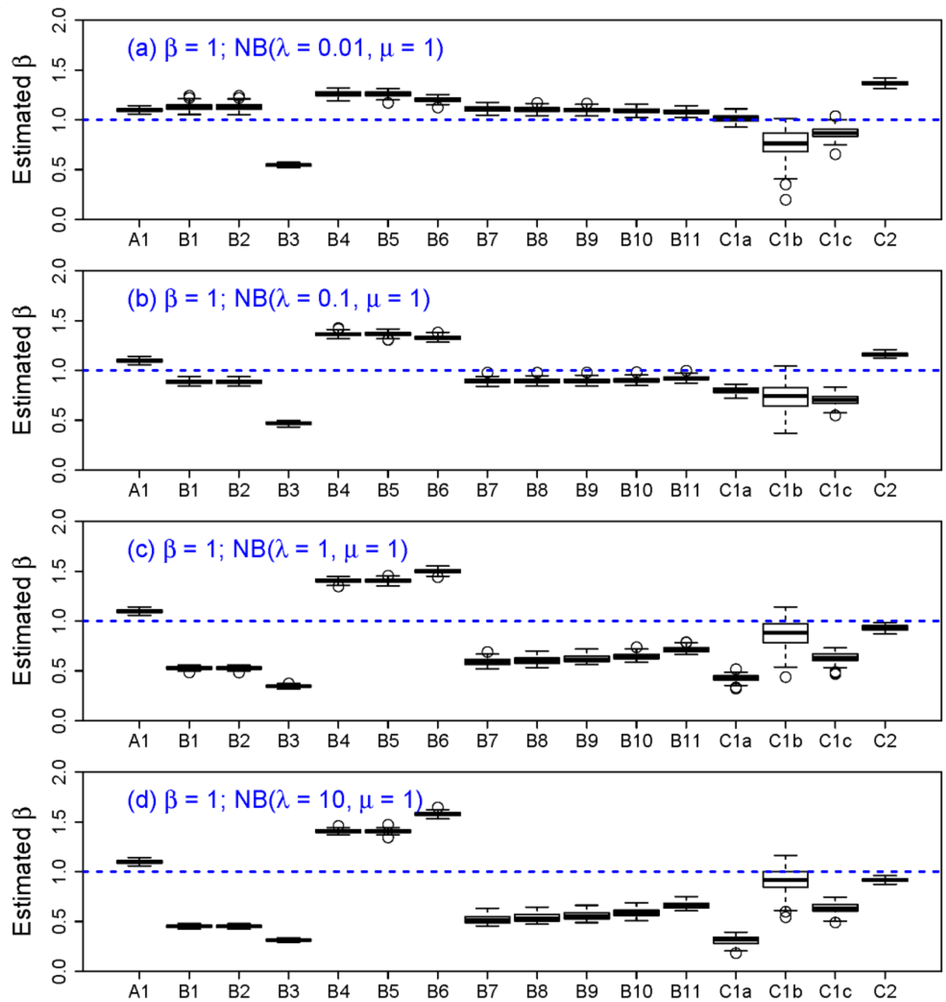

Figure 5. Comparison of bias in estimated spectral slope in irregular data that are simulated with prescribed $\beta = 1$ (~~30~~100 replicates), series length of 9125, and gap intervals simulated with (a) NB ($\lambda = 0.01$, $\mu = 1$), (b) NB ($\lambda = 0.1$, $\mu = 1$), (c) NB ($\lambda = 1$, $\mu = 1$), and (d) NB ($\lambda = 10$, $\mu = 1$). The blue dashed lines indicate the true β value.

Commented [QZ1]: Updated figure with 100 replicate runs

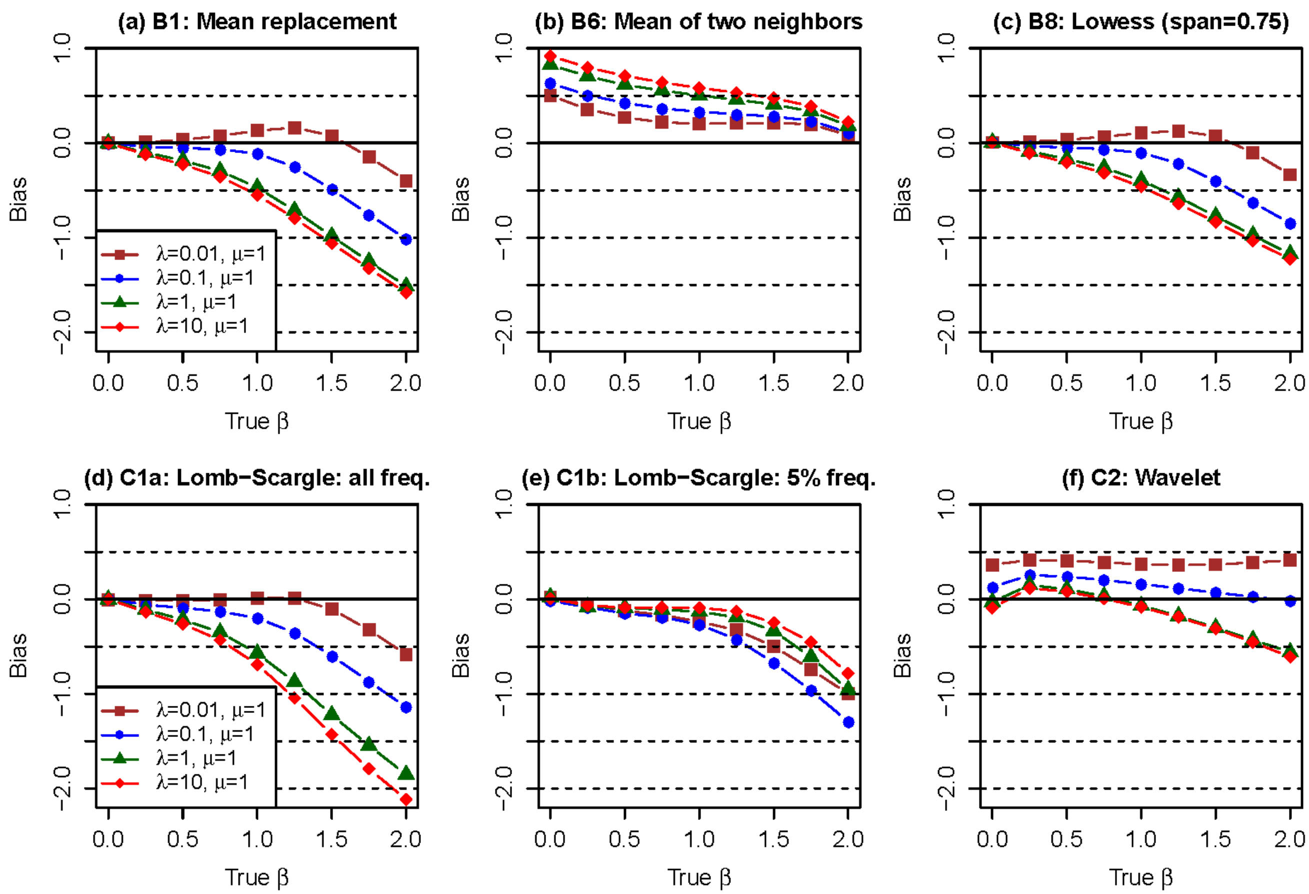

Figure 6. Comparison of bias in estimated spectral slope in irregular data that are simulated with varying prescribed β values (~~30~~100 replicates), series length of 9125, and mean gap interval of 2 (*i.e.*, μ = 1).

Commented [QZ2]: Updated figure with 100 replicate runs

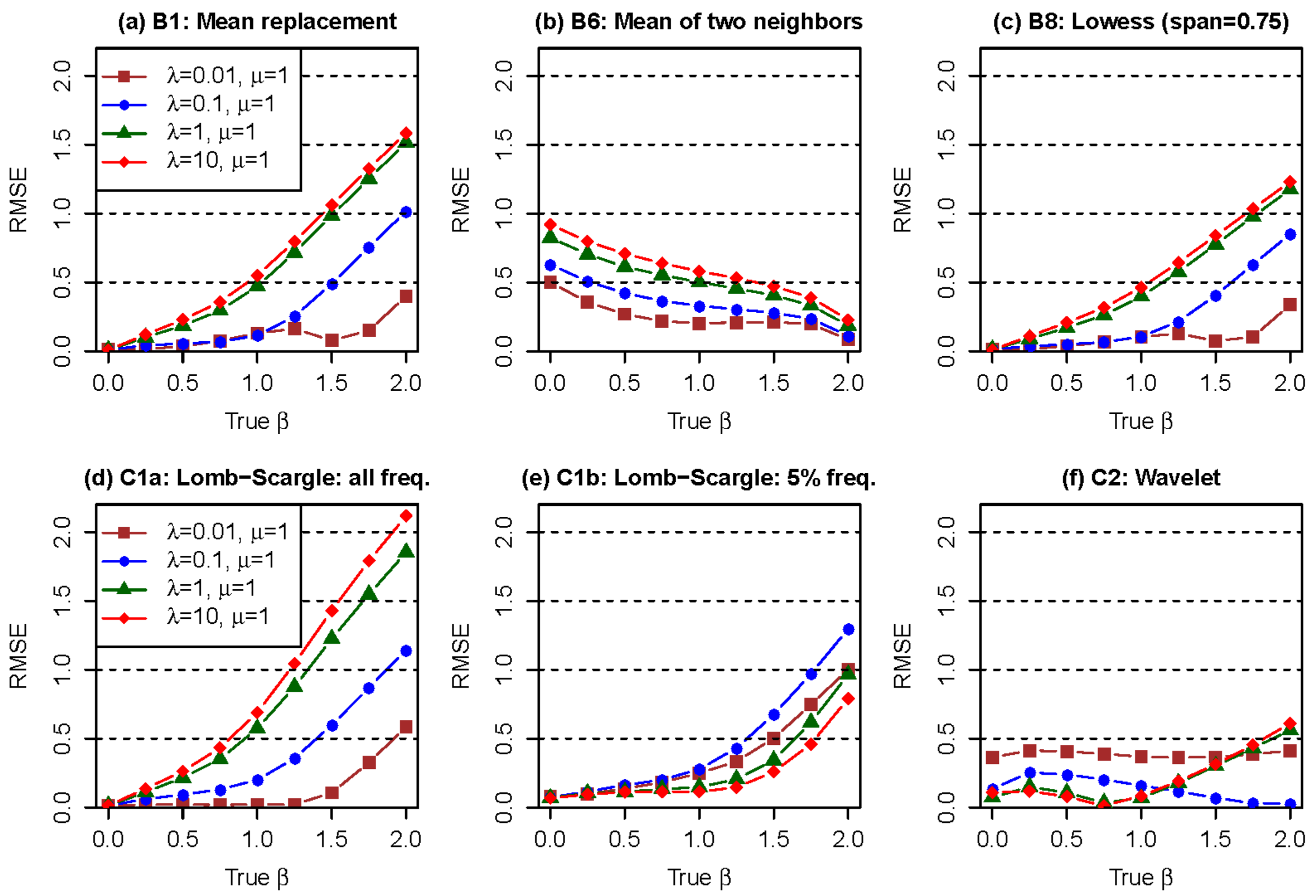

Figure 7. Comparison of root-mean-squared error (RMSE) in estimated spectral slope in irregular data that are simulated with varying prescribed β values (~~30~~100 replicates), series length of 9125, and mean gap interval of 2 (*i.e.*, $\mu = 1$).

Commented [QZ3]: Updated figure with 100 replicate runs

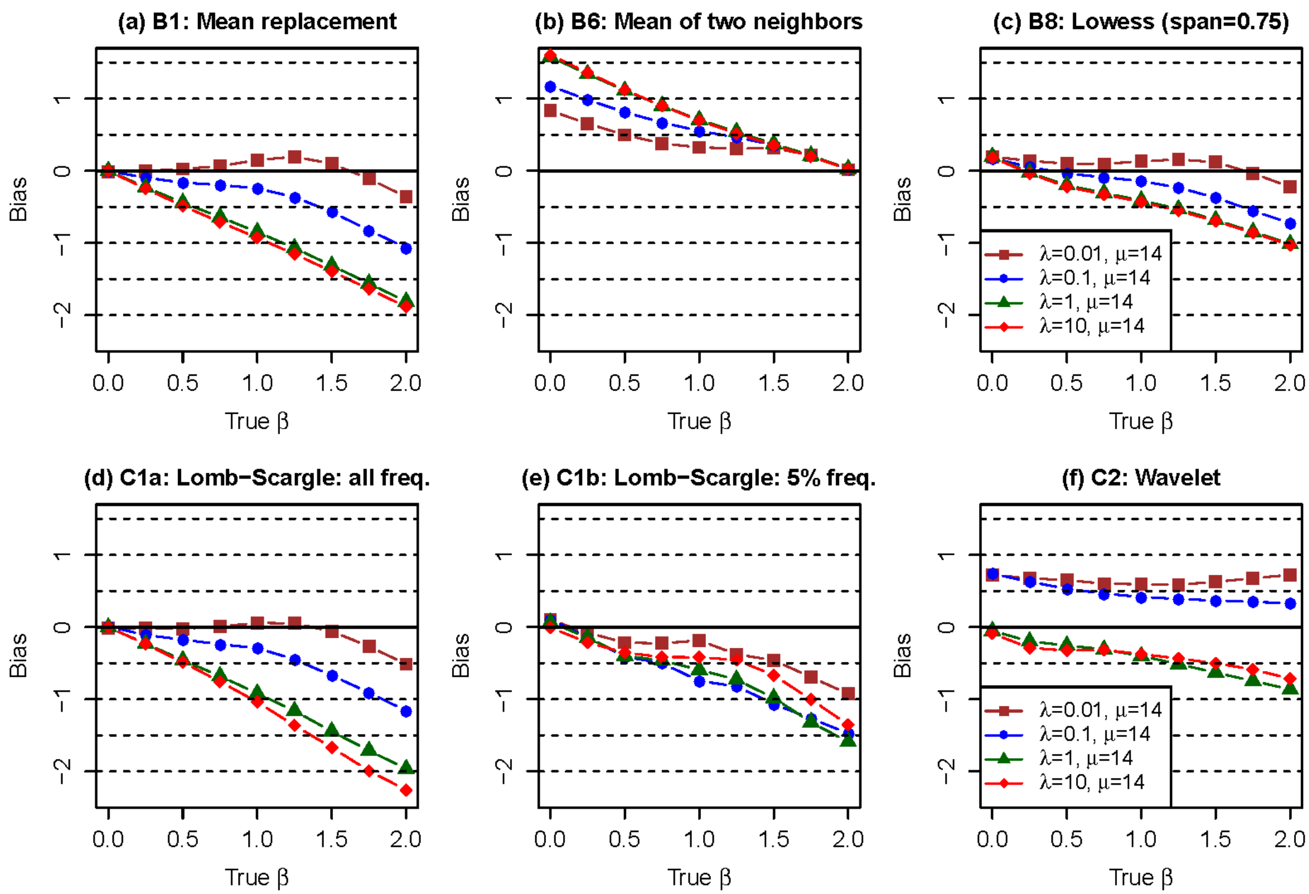

Figure 8. Comparison of bias in estimated spectral slope in irregular data that are simulated with varying prescribed β values (~~30~~100 replicates), series length of 9125, and mean gap interval of 15 (*i.e.*, μ = 14).

Commented [QZ4]: Updated figure with 100 replicate runs

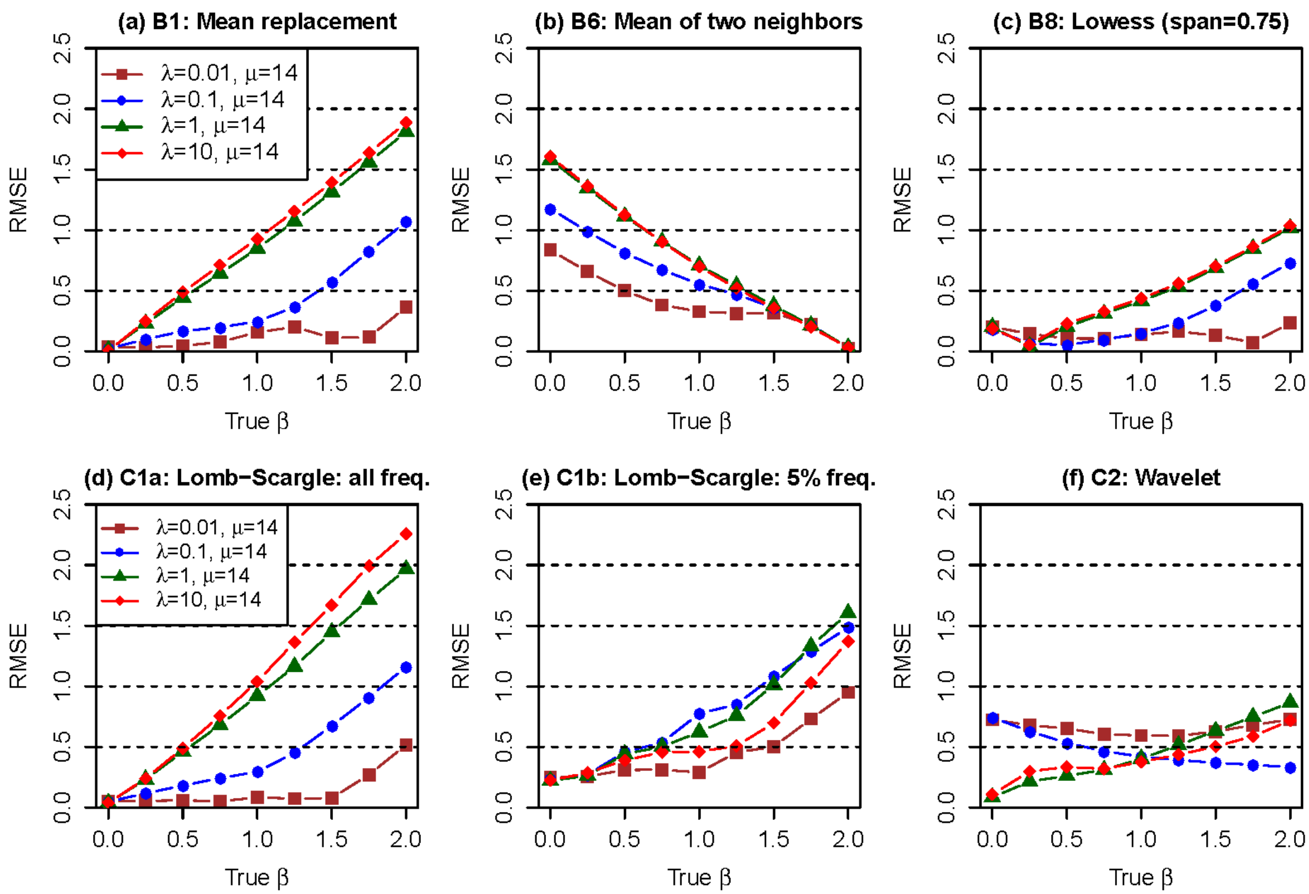

Figure 9. Comparison of root-mean-squared error (RMSE) in estimated spectral slope in irregular data that are simulated with varying prescribed β values (~~30~~100 replicates), series length of 9125, and mean gap interval of 15 (*i.e.*, μ = 14).

Commented [QZ5]: Updated figure with 100 replicate runs

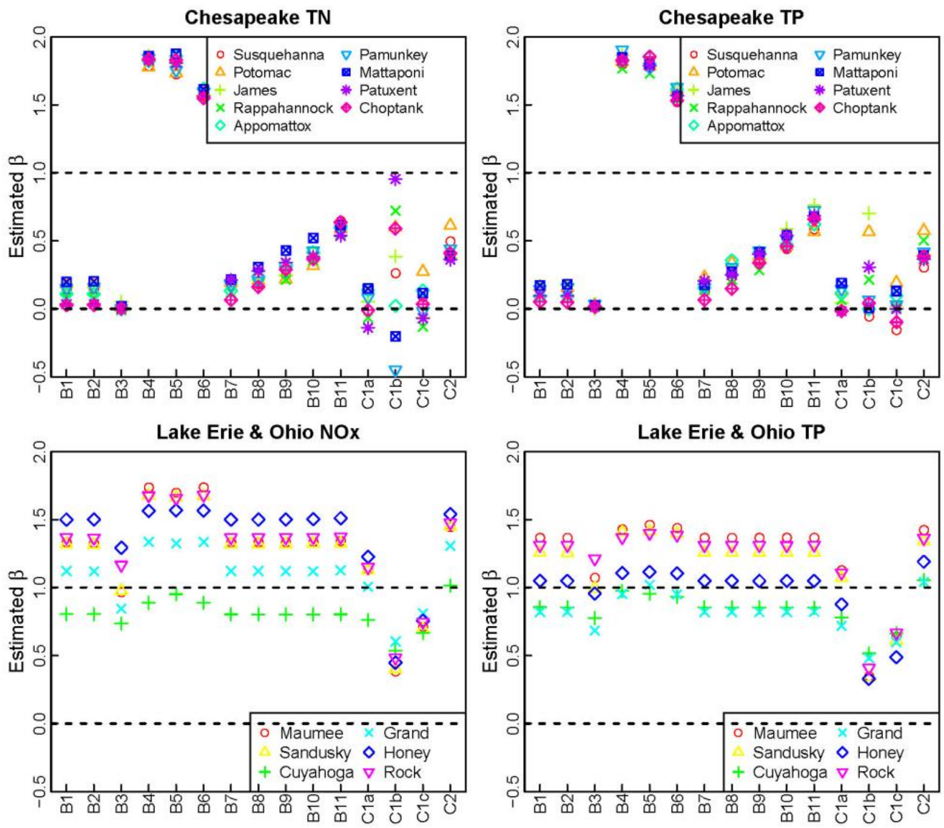

Figure 10. Quantification of spectral slope in real water-quality data from the two regional monitoring networks, as estimated using the set of examined methods. All estimations were performed on concentration residuals (in natural log concentration units) after accounting for effects of time, discharge, and season. The two dashed lines in each panel indicate white noise (β = 0) and pink (flicker) noise (β = 1), respectively. See **Table 1** for site and data details.