Geophysical Research Letters

RESEARCH LETTER
10.1029/2022GL101177

# Reconstruction of Sparse Stream Flow and Concentration Time-Series Through Compressed Sensing

**Kun Zhang[1,2], Wasif Bin Mamoon[1], E Schwartz[3], and Anthony J. Parolari[1]**

[1]Department of Civil Construction, and Environmental Engineering, Marquette University, Milwaukee, WI, USA, [2]Department of Civil and Environmental Engineering, Seattle University, Seattle, WA, USA, [3]Department of Environmental Resources Engineering, SUNY ESF, Syracuse, NY, USA

**Key Points:**

- Streamflow and concentration signals are characterized as sparse in Fourier frequency domain
- Compressed sensing (CS) effectively reconstructs 15-min-scale stream flow and concentration signals with only 5%–10% of measurements needed
- CS well estimates stream $NO_x$-N and total phosphorus loads with an effective sampling frequency of 10 and 0.4 days, respectively

**Supporting Information:**

Supporting Information may be found in the online version of this article.

**Correspondence to:**

K. Zhang,
kun.zhang@marquette.edu

**Citation:**

Zhang, K., Bin Mamoon, W., Schwartz, E, & Parolari, A. J. (2023). Reconstruction of sparse stream flow and concentration time-series through compressed sensing. *Geophysical Research Letters*, *50*, e2022GL101177. https://doi.org/10.1029/2022GL101177

Received 12 SEP 2022
Accepted 14 DEC 2022

**Author Contributions:**

**Conceptualization:** Anthony J. Parolari
**Data curation:** Kun Zhang, Wasif Bin Mamoon
**Formal analysis:** Kun Zhang
**Funding acquisition:** Anthony J. Parolari
**Investigation:** E Schwartz
**Methodology:** E Schwartz
**Project Administration:** Anthony J. Parolari
**Resources:** Wasif Bin Mamoon, E Schwartz
**Software:** Kun Zhang
**Supervision:** Anthony J. Parolari
**Validation:** Kun Zhang



**Abstract** Monitoring water quality at high frequency is challenging and costly. Compressed sensing (CS) offers an approach to reconstruct high-frequency water quality data from limited measurements, given that water quality signals are commonly "sparse" in the frequency domain. In this study, we investigated the sparsity of stream flow and concentration time-series and tested reconstruction with CS. All stream signals were sparse using 15-min discrete time-series transformed to the Fourier domain. Stream temperature, conductance, dissolved oxygen, and nitrate plus nitrite ($NO_x$-N) concentration were sparser than discharge, turbidity, and total phosphorus (TP) concentration. CS effectively reconstructed these signals with only 5%–10% of measurements needed. Stream $NO_x$-N and TP loads were well estimated with errors of $-6.6\% \pm 3.8\%$ and $-9.0\% \pm 2.9\%$ with effective sampling frequencies of 10 and 0.4 days, respectively. For broader applications in environmental geosciences and engineering domains, CS can be integrated with dimensionality reduction and optimization techniques for more efficient sampling schemes.

**Plain Language Summary** Most natural environmental signals can be represented by less numbers or compressed. Signals that can be compressed can be accurately monitored with many fewer measurements than currently used in practice. This study reports and compares the ability of streamflow and water quality signals to be compressed. The results demonstrate that stream flow and several water quality metrics can be compressed when expressed in a specific mathematical format, called the frequency space. Electrical conductivity, dissolved oxygen, and nitrate plus nitrite concentration were found to be more compressible than stream flow, turbidity, and total phosphorus concentration. These variables can be effectively estimated with only 5%–10% of the number of measurements currently used in practice. Therefore, the compressed sensing method has potential to achieve large reductions in sampling effort when collecting data for environmental science and engineering projects, with associated time and cost savings.

## 1. Introduction

Streamflow and water quality monitoring data are needed to characterize stream physical and ecological regimes, responses to climate and land use change, impairments, and pollutant loads for regulatory oversight and management (Bremer et al., 2020; Burns et al., 2019; Crawford et al., 2015; Pellerin et al., 2014; Pluth et al., 2021; Reynolds et al., 2016; Skeffington et al., 2015; Smith et al., 1997). Many studies agree that high-frequency or even near-continuous sampling is necessary to capture the dynamics of flow and concentration and to effectively estimate pollutant loads (Cassidy & Jordan, 2011; Gao et al., 2020; Jones et al., 2012; Kerr et al., 2018; Minaudo et al., 2017; Pellerin et al., 2014; Reynolds et al., 2016; Skeffington et al., 2015; Thompson et al., 2021). However, despite recent advances in sensor technology (Rode et al., 2016), high-frequency sampling remains prohibitive and stream assessment still relies heavily on infrequent grab sampling, especially for emerging contaminants.

Many environmental signals or time-series (e.g., soil moisture and temperature) are "sparse," meaning that these signals can be characterized in an appropriate transform domain with a number of coefficients less than the number of sampled times (Katul et al., 2007; Kirchner & Neal, 2013; Parolari et al., 2021; Vasseur & Yodzis, 2004). Compressed Sensing (CS), a signal processing technique also known as compressive sensing (or sampling), leverages the inherent sparsity of a signal or process to reconstruct high resolution data from a relatively small number of samples (Donoho, 2006). For example, Figure 1a illustrates a synthetic signal with two frequencies. When transformed to a Fourier frequency space, the signal can be represented by only two active coefficients at

**Visualization:** Kun Zhang
**Writing – original draft:** Kun Zhang, E Schwartz
**Writing – review & editing:** Kun Zhang, Wasif Bin Mamoon, Anthony J. Parolari

these two frequencies (Figure 1b). With very few samples taken (Figure 1a), the signal can be well reconstructed (Figure 1d).

CS has the potential to substantially reduce the number of samples required to estimate high-frequency stream water quality data and corresponding loads. CS has been demonstrated to be effective in signal compression and recovery in various contexts including photographs, MRI images (Lustig & Donoho, 2008), radar data (Zhang et al., 2010), and land cover data (Wei et al., 2017). Based on the known sparsity of water quality parameter time-series, we hypothesized that high-frequency water quality data can be obtained from a low-frequency monitoring strategy using CS.

To test whether CS can be used to reduce water quality sampling requirements, stream flow and concentration time-series were collected at 45 stream gauges in the Midwest United States. Using this data, we first demonstrate the sparsity of daily and 15-min stream flow and concentration time-series, which included discharge, temperature, specific conductance (SC), turbidity, dissolved oxygen (DO), nitrate plus nitrite as nitrogen concentration, and total phosphorus (TP) concentration. Then, CS was applied to reconstruct these time-series and estimate the annual loads for nitrate plus nitrite and TP based on different numbers of samples, corresponding to different "effective" sampling intervals.

## 2. Data and Model

### 2.1. Signal Sparsity and Compressed Sensing

Flow or concentration signals are collected as discrete time-series $\boldsymbol{x}$ of length $n$, which can be represented by a linear combination of vectors $\boldsymbol{s} \in R^n$ in a specific basis $\boldsymbol{\Psi} \in R^{n\times n}$,

$$\boldsymbol{x} = \boldsymbol{\Psi}\boldsymbol{s} \tag{1}$$

Similar to many other natural signals, flow and concentration signals are normally sparse in a certain basis, meaning that the time-series can be represented only by a few active modes with large mode amplitudes when the discrete time-series $\boldsymbol{x}$ is represented in an appropriate coordinate system or basis (Figures 1a and 1b).

"Lorenz" (after the economist Max Lorenz) curves (LCs) are potential tools that can represent the "sparsity" of signals. The LC plots the cumulative variance of a signal as a function of the number of coefficients included in the sum, sorted from smallest to largest (Katul et al., 2007) (Figure 1c). In the time domain, the coefficients that contribute to the variance are the squared deviations from the average, that is, $\left(\boldsymbol{x}_t - \overline{\boldsymbol{x}}\right)^2$, for all time instances. In the frequency domain, the coefficients are the power spectral density estimated from the Fourier transform, that is, $\boldsymbol{S}_f$. A time-series is considered sparse in the corresponding domain if a small percentage of coefficients explains the majority of variance (i.e., LC far from the 1:1 line as shown in Figure 1c). In contrast, a time-series is not sparse if most of the coefficients contribute equally to the variance (Katul et al., 2007). Here, sparsity was quantified as the proportion of coefficients associated with a unit loss in energy (or a slope of 1 in LC), $K_{\text{us}}$. The value of $K_{\text{us}}$ ranges between 0 and 1 where a value close to 0 indicates the LC is nearly perfectly balanced in terms of energy among the coefficients and a value close to 1 implies the LC is energetically imbalanced as the variability is contained in only a few coefficients. In other words, a higher value of $K_{\text{us}}$ indicates a sparser signal.

With a subset of measurements $\boldsymbol{y}$ in the flow or concentration time-series $\boldsymbol{x}$ determined by the measurement matrix $\boldsymbol{C}$ (Equation 2; Figure 1a), if the sparsest $\boldsymbol{s}$, denoted as $\hat{\boldsymbol{s}}$ that has the largest number of zero coefficients, can be found, the original time-series $\boldsymbol{x}$ can be reconstructed through inversion of Equation 1 (Figure 1d).

$$\boldsymbol{y} = \boldsymbol{C}\boldsymbol{x} = \boldsymbol{C}\boldsymbol{\Psi}\boldsymbol{s} \tag{2}$$

Equation 2 is an underdetermined system of equations for $\boldsymbol{s}$, meaning that there are infinite number of solutions for $\boldsymbol{s}$ since the number of equations is much smaller than the number of entries of $\boldsymbol{s}$. The solution for $\hat{\boldsymbol{s}}$ corresponds to pursuing the smallest $\ell_0$-norm. However, solving the $\ell_0$-norm problem directly is intractable, and, thus, a convex $\ell_1$-minimization problem was solved to reduce the complexity (Equation 3) (Candès et al., 2006; Donoho, 2006),

$$\hat{\boldsymbol{s}} = \underset{\boldsymbol{s}'}{\operatorname{argmin}} \|\boldsymbol{s}'\|_1, \text{ such that } \boldsymbol{y} = \boldsymbol{C}\boldsymbol{\Psi}\boldsymbol{s}' \tag{3}$$

To ensure that the sparsest $\boldsymbol{s}$ with respect to $\boldsymbol{\Psi}$ can be found, the samples taken should represent a wide range of the temporal basis functions in $\boldsymbol{\Psi}$. In other words, the sampling matrix $\boldsymbol{C}$, which determines when the samples

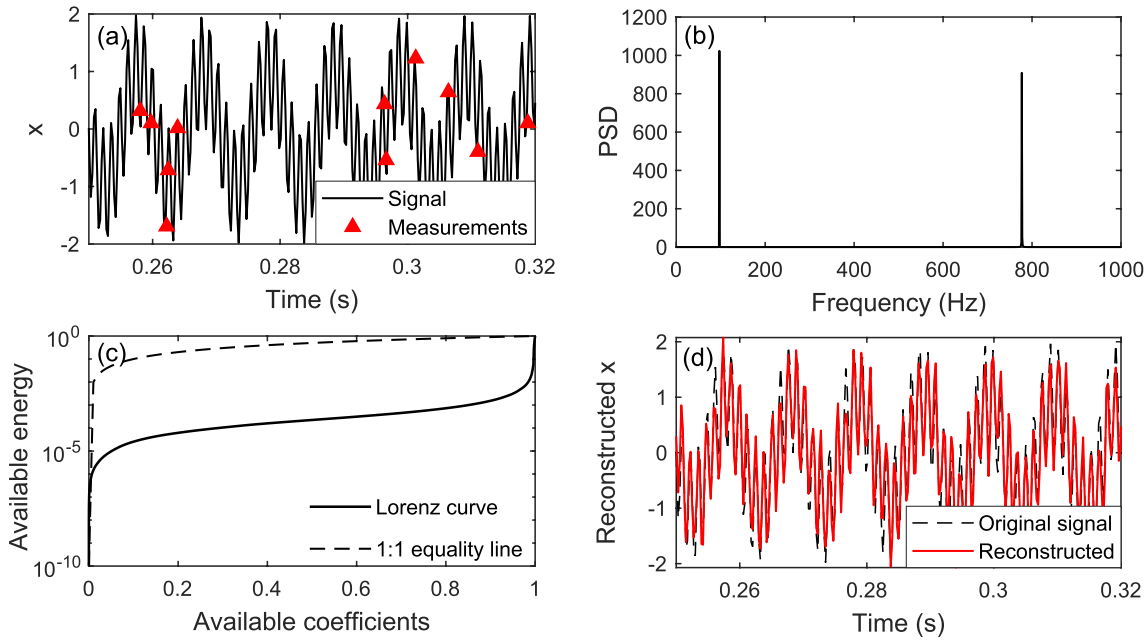


**Figure 1.** Representation of a sparse signal by Lorenz curve and its reconstruction using compressed sensing (CS). (a) Full time-series with two active frequencies. (b) Power spectral density of the signal in Fourier basis. (c) Lorenz curve of the signal. (d) Reconstructed signal using CS. Modified from Brunton and Kutz (2022).

are taken, should be "incoherent" to the representation basis $\mathbf{\Psi}$ (Candes & Wakin, 2008). Uniform random sampling in time is normally the best way to ensure maximum incoherence.

### 2.2. Stream Monitoring Data and Analysis

Stream flow and concentration time-series were retrieved from the United States Geologic Survey (USGS) National Water Information System (https://waterdata.usgs.gov/nwis). Forty-five stream gauges were identified for testing based on the high-frequency availability of TP and nitrate plus nitrite as nitrogen ($NO_x$-N) concentrations. Daily-scale stream flow and concentration data from 2015 to 2021 was retrieved for these 45 gauges. The data set included discharge ($Q$; $n = 44$), temperature (Temp; $n = 25$), specific conductance (SC; $n = 8$), turbidity (Turb; $n = 18$), dissolved oxygen (DO; $n = 8$), nitrate plus nitrite as nitrogen concentration ($NO_x$-N; $n = 20$), and TP concentration (TP; $n = 27$) with the record length and completeness varying among gauges. All gauges were located in the Midwest region of the United States (Figure S1 in Supporting Information S1). In addition to daily-scale data at all 45 gauges, one 15-min time-series during the period (2015–2021) with the most complete data set was retrieved for each variable. The gauge IDs, variables, and record lengths for the daily-scale and 15-min data sets can be found in Tables S1 and S2 in Supporting Information S1, respectively.

LCs in both the time and Fourier frequency domains were generated to determine and compare the level of sparsity of stream flow and concentration signals. They were generated from daily stream flow and concentration monitoring data in all available gauges. In addition, CS was utilized to reconstruct daily and 15-min stream flow and concentration time-series. Only one gauge for each variable was selected for reconstruction. The gauge information can be found in Table S2 in Supporting Information S1. To increase the reliability of the results and reduce the potential impact of the randomly determined sampling times on the results, the CS reconstruction was repeated 10 times for each gauge and each variable. CS was tested with different number of samples taken, ranging from 0.5% to 50% of total length of time-series. The samples were taken at random, uniformly distributed times to ensure that the sparsest $\boldsymbol{s}$ with respect to $\mathbf{\Psi}$ can be found and the time-series can be reconstructed using less samples and with high probability (Candes & Wakin, 2008). The compressive sampling matching pursuit (CoSaMP) algorithm (Needell & Tropp, 2009) was utilized to solve the sparsest solution of Equation 3. Nash Sutcliffe Efficiency (NSE) (Nash & Sutcliffe, 1970), which determines the relative magnitude of the residual variance, $\left(\boldsymbol{x}_o^t - \boldsymbol{x}_m^t\right)^2$, compared to the measured data variance, $\left(\boldsymbol{x}_o^t - \overline{\boldsymbol{x}_o}\right)^2$, was used to quantify the goodness of fit between the measured and reconstructed series (Equation 4).

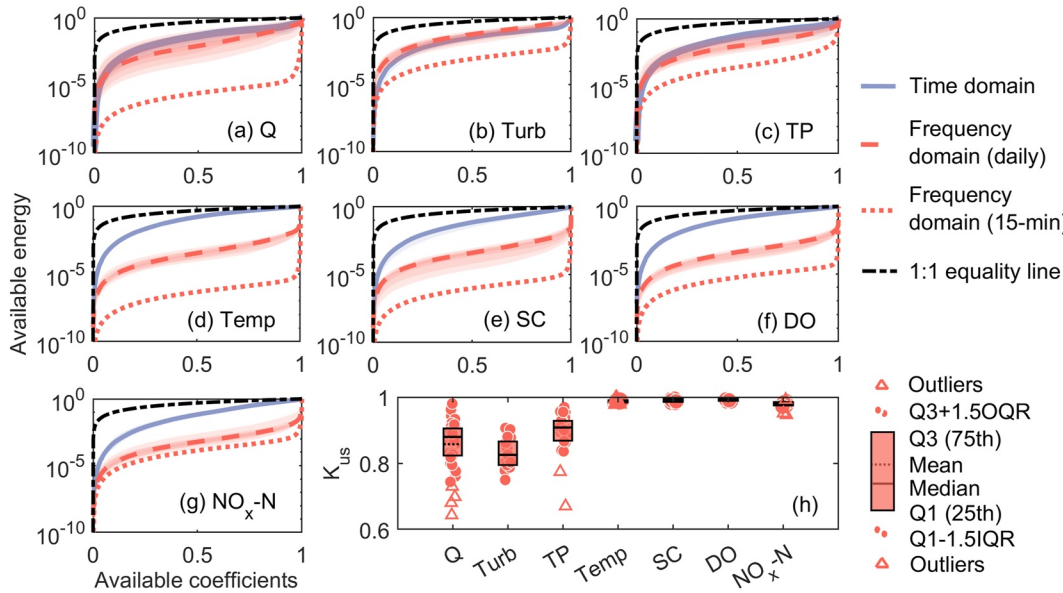


**Figure 2.** Sparsity of daily and 15-min stream flow and concentration time-series in the time and Fourier frequency domains: (a) $Q$, (b) Turb, (c) total phosphorus (TP), (d) Temp, (e) specific conductance (SC), (f) dissolved oxygen (DO), and (g) $NO_x$-N respectively refer to discharge, turbidity, TP concentration, SC, DO, and nitrate plus nitrite as nitrogen concentration. The shaded regions refer to the 25th, 50th, 75th, and 90th percentiles across gauges. (h) The proportion of coefficients that correspond to the point at which a unit loss in a coefficient results in a unit loss in energy in the Fourier frequency domain ($K_{us}$). In panel (h), $Q1$ and $Q3$ refer to the 25th and 75th percentiles of the data. IQR refers to interquartile range (Q3-Q1). The outliers are values that are located outside 1.5 times IQR above $Q3$ or below $Q1$.

$$NSE = 1 - \frac{\sum_{t=1}^{N} \left(x_o^t - x_m^t\right)^2}{\sum_{t=1}^{N} \left(x_o^t - \overline{x_o}\right)^2} \tag{4}$$

## 3. Results

### 3.1. Sparsity of Stream Flow and Concentration Time-Series

Stream flow and concentration signals are sparse in the Fourier domain. Based on the daily-scale time-series data, temperature (Temp), SC, DO, and nitrate plus nitrite as nitrogen concentration ($NO_x$-N) were substantially sparser than discharge ($Q$), turbidity (Turb), and total phosphorus concentration (TP) (Figure 2). The $K_{us}$ values obtained imply that the variances of $Q$, Turb, and TP can be represented by anywhere from 2% to 36% of the most energetic coefficients in the Fourier frequency domain. On the other hand, the variances of Temp, SC, DO, and $NO_x$-N can be well represented with only 0.3%–5.3% most energetic coefficients (Figure 2h).

The true signal sparsity can be better characterized with time-series at finer temporal resolution. Using 15-min data, the variances of all seven time-series can be represented by only 0.1%–2.1% of the most energetic coefficients in the Fourier frequency domain (Figure 2). We presume that the 15-min data provides a more faithful representation of the "true" process dynamics than the daily data. Therefore, it is reasonable to expect that all variables analyzed here are truly sparse, although the daily time-series are less able to reveal this underlying sparsity.

### 3.2. Stream Data Reconstructions With Compressed Sensing

The stream flow and concentration time-series were effectively reconstructed using compressed sensing. The goodness of fit for the reconstruction is illustrated in Figure 3 using 15-min data with 10% of measurements as an example (NSE = 0.895–0.997 for all different variables). The peak values of Turb and TP during extreme storm events were underestimated (Figures 3b-1, 3b-2, 3c-1, and 3c-2), resulting in a relatively smaller goodness of fit.

Comparatively, sparser time-series were more effectively reconstructed based on fewer measurements than less sparse time-series. At the 15-min resolution, all the time-series were reconstructed nearly perfectly (NSE ≈ 1)

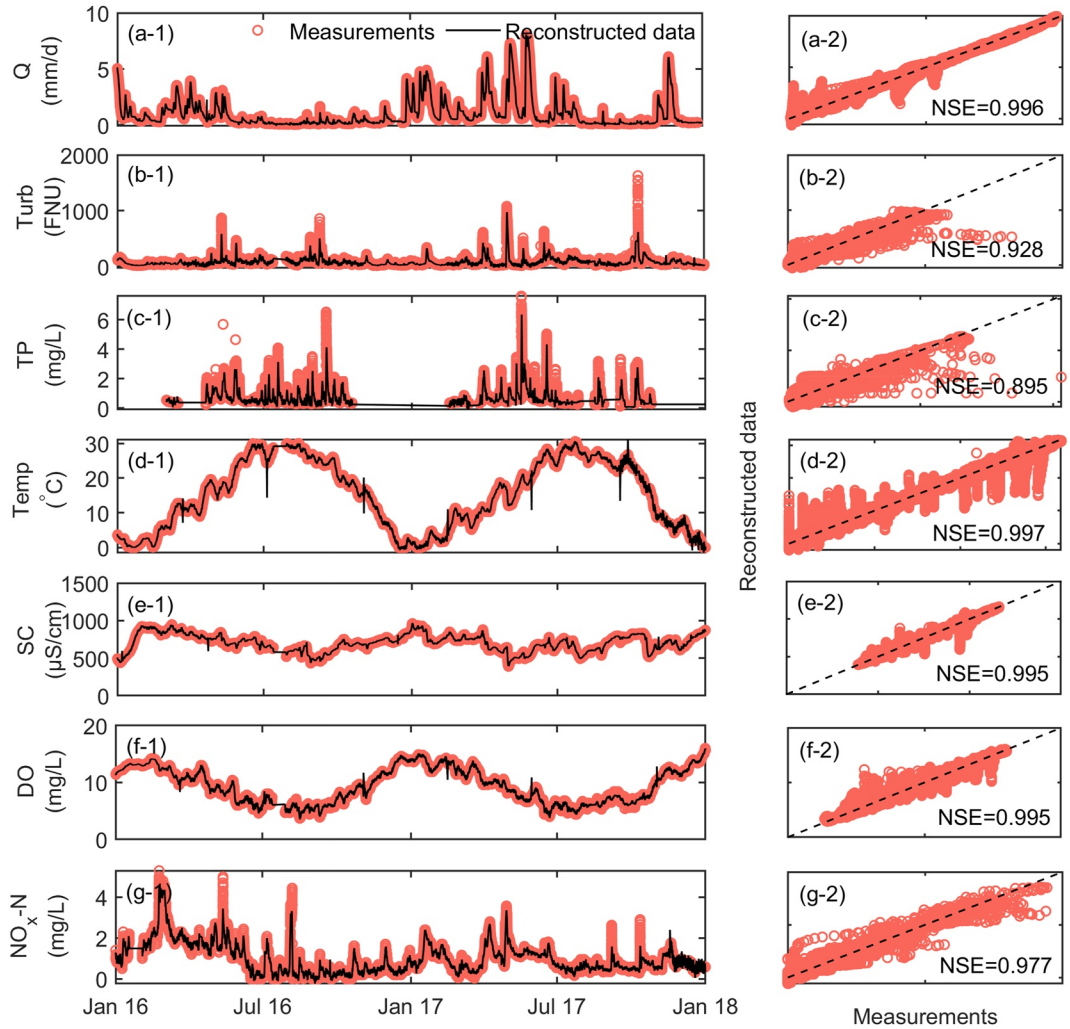


**Figure 3.** Comparison of 15-min resolution measured and reconstructed time-series (left) with corresponding goodness of fit (right): (a-1 and a-2) discharge (*Q*), (b-1 and b-2) turbidity (Turb), (c-1 and c-2) total phosphorus concentration, (d-1 and d-2) temperature (Temp), (e-1 and e-2) specific conductance, (f-1 and f-2) dissolved oxygen, and (g-1 and g-2) nitrate plus nitrite as nitrogen concentration ($NO_x$-N) with measurement percentage of 10%.

with 5%–10% of measurements. Only 10% of measurements were needed for the less sparse *Q*, Turb, and TP, while only 5% of measurements were needed for the sparser Temp, SC, DO, and $NO_x$-N (Figure 4). At the daily resolution, less sparse variables (*Q*, Turb, and TP) were not well-reconstructed even with half of the measurements. On the other hand, for sparser variables (Temp, SC, and DO), the full time-series was reconstructed nearly perfectly (NSE ≈ 1) with ~10% of daily measurements, equivalent to ~0.1% of 15-min measurements. Comparatively, the reconstruction effectiveness for $NO_x$-N at the daily scale was slightly lower than Temp, SC, and DO, despite its high sparsity level (Figure 4).

The samples taken for reconstruction were randomly and uniformly distributed throughout the period as shown in Figures S2–S3 in Supporting Information S1. Although this random, irregular sampling scheme can be difficult to implement in practice when grab sampling is used, it can potentially support field sampling campaigns with sampling constrained to weekdays or during daylight times. Results demonstrated that taking samples only on weekdays or during daylight times does not significantly affect the reconstruction efficiency (Figure S5 in Supporting Information S1). However, when samples are restricted to only daylight hours on weekdays, the performance of CS for some variables was diminished (i.e., turbidity, temperature, specific conductance, and DO), but was not affected for others (i.e., discharge, TP, and $NO_x$-N). It should be noted that the reconstruction effectiveness using CS varied when performed on time-series of different lengths (Figure S4 in Supporting

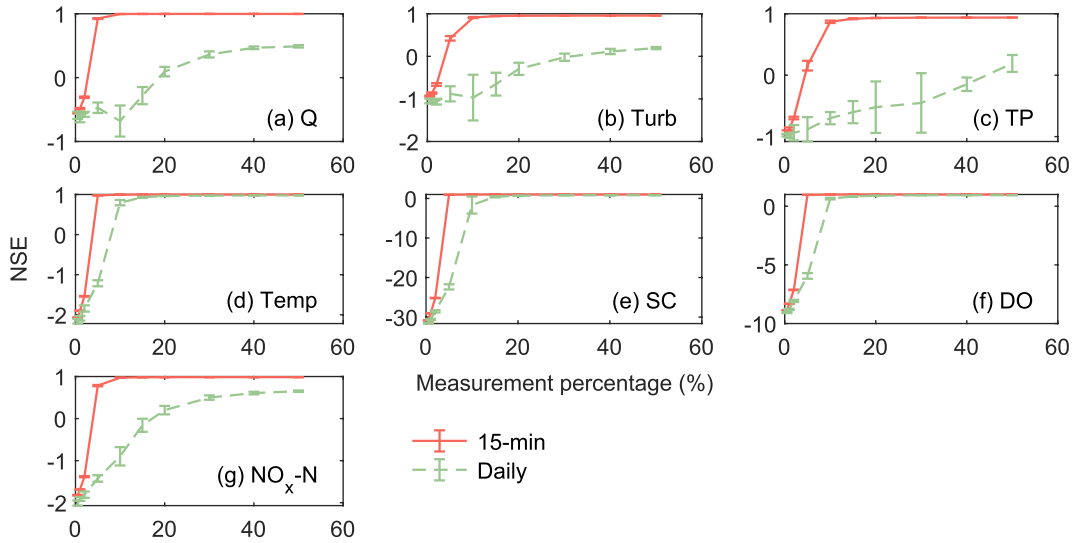


**Figure 4.** (a–f) Variation of Nash Sutcliffe Efficiency (NSE) with data compression percentage using compressed sensing (CS). (a) *Q*, (b) Turb, (c) *P*, (d) Temp, (e) specific conductance (SC), (f) dissolved oxygen (DO), and (g) $NO_x$-N respectively refer to flow discharge, turbidity, total phosphorus concentration, temperature, SC, DO, and nitrate plus nitrite as nitrogen concentration. The error bars represent the standard deviations of NSE values around the medians for 10 iterations of CS reconstruction.

Information S1). Longer or higher-resolution time-series are more representative of the true signal and, therefore, the samples collected incorporate a wider range of hydro-environmental conditions.

### 3.3. Stream Nutrient Load Estimation With Compressed Sensing

The annual nutrient load can be more accurately estimated when sampling at a higher frequency and when using CS to reconstruct at a higher frequency. Testing different effective sampling frequencies and reconstruction frequencies provided insight into application of CS to reduce the sample requirements for annual load estimation (Figure 5). Using the 15-min data as the true signal, the true annual $NO_x$-N load was estimated with an error of $-6.6\% \pm 3.8\%$ using CS with an effective sampling frequency of ~10 days projected to a daily time-series. The true annual TP load was estimated with an error of $-9.0\% \pm 2.9\%$ using CS with an effective sampling frequency of ~0.4 days projected to a 1-hr time-series. Note that if the benchmark load is defined as that obtained with daily samples (as is common in practice), the effective sampling frequencies with CS can be even lower. The daily TP load can be well estimated with samples every 1.7 days. However, these estimates are still biased relative to the "true" load, estimated here from the 15-min data. In all cases, the optimal sampling frequency for $NO_x$-N was lower than that for TP because the $NO_x$-N signal is sparser in the frequency domain (as observed in this study).

## 4. Discussion and Conclusions

The sparsity of different stream flow and concentration signals in the frequency domain varied between one another because of their unique time-frequency characteristics which could be related to their connectivity to climatic cycles and watershed geophysical conditions. In general, more process-driven signals are likely to have fewer active modes in the frequency domain, or sparser, and therefore more easily reconstructed. This can include signals dominated by diel and seasonal energy cycles, such as temperature, SC, DO, and $NO_x$-N (Gammons et al., 2011), or signals dominated by storage, such as snowmelt- or groundwater-dominated discharge. On the other hand, more noise-driven signals possess more randomness and are "flashier." Thus, these signals are likely to have more active modes in the frequency domain, and, thus, are less sparse and require more samples to be reconstructed. This can include signals dominated by precipitation variability, such as the discharge, turbidity, and TP concentration signals evaluated in this study (Banner et al., 2009; Zabaleta et al., 2007). Due to different levels of dependence on hydrologic and geophysical processes, the sparsity of hydro-environmental signals can bear regional signatures and show spatial patterns.

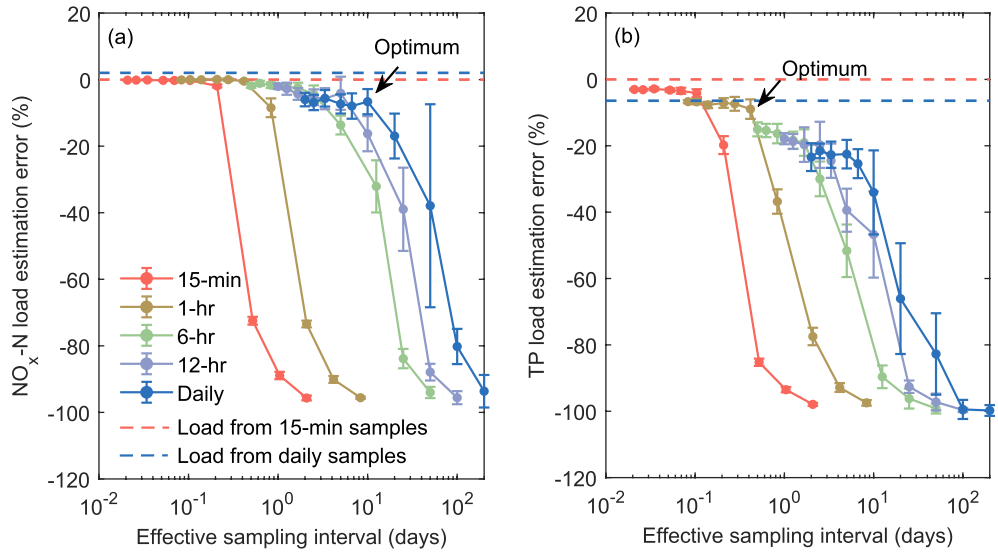


**Figure 5.** Variation of annual load estimation errors for (a) nitrate plus nitrite as nitrogen ($NO_x$-N) and (b) total phosphorus with effective sampling frequencies using Compressed sensing (CS). Effective sampling frequencies ranged between those corresponding to 0.5% and 50% of the data points for each time-series. The dashed lines represent the annual load estimated from uniformly subsampled 15-min data (red), as the "true" load, and daily samples (blue). The error bars represent the standard deviations of annual load estimation error around the medians for 10 iterations of CS reconstruction.

Several methods have been previously reported to increase sampling efficiency and reduce pollutant load estimation errors. These approaches include high-frequency sampling, random subset generation, automated sampling based on flow, complementing low-frequency fixed-interval sampling with higher frequencies during storm events, and data-model integration (Cassidy & Jordan, 2011; Jones et al., 2012; Minaudo et al., 2017; Reynolds et al., 2016; Rozemeijer et al., 2010; Thompson et al., 2021). In these cases, estimated biases ranged from 1% to 104%, with the best performance achieved by combining high-frequency storm sampling with models (Minaudo et al., 2017; Rozemeijer et al., 2010). Here, when using the 15-min data as the "true" benchmark, CS with an effective sampling frequency of ∼10 and ∼0.4 days achieved a good estimation of $NO_x$-N and TP loads, respectively. The estimation error comes from both the error in CS reconstruction (due to missing frequencies in the samples) and the bias embedded in the time-series (due to missing frequencies in the underlying data). Additional work is needed to quantify the trade-off between sampling costs and load estimation accuracy across all available methods, and/or to correct the bias embedded in the samples taken and further improve the load estimation accuracy.

In addition to achieving small bias in annual load estimation, CS was able to reconstruct accurate high-frequency stream data from low-frequency samples without requiring any prior knowledge of the relevant processes. In this application, we were able to achieve a 90%–95% reduction in the number of samples needed to estimate 15-min stream flow and concentration time-series data. For the sparsest variables (temperature, conductance, and DO) which show strong diel and/or seasonal fluctuations, daily data was effectively reconstructed with 90% reduction in the number of samples needed. This corresponds to a 5–10-day effective sampling interval, representing a measurable reduction in sampling effort.

However, CS has some limitations and there are opportunities for improvement for practical applications. First, although reducing the number of measurements needed can save sensor and battery life and reduce data storage and transmission for monitoring networks, it is difficult to replace regularly scheduled weekday/daylight grab sampling with varied-interval sampling as required here. Thus, in future work, CS could be performed on measurements retrieved from other sources such as remotely sensed data. Second, CS may not work well for signals not sufficiently sparse in the frequency domain. Thus, instead of projecting the measurements onto the frequency domain, dimensionality reduction techniques such as singular value decomposition can be utilized to identify an appropriate minimal space where the signals can be sparsely represented (Chatterjee, 2000). Third, the reconstruction effectiveness using CS depends on the amount of information contained in the samples taken. Random-interval sampling as required by CS is not optimal to maximize the information content. In future work,

CS can be integrated with techniques to optimize the sampling times to ensure best reconstruction with minimal samples needed (Manohar et al., 2018; Ohmer et al., 2022).

## Data Availability Statement

All the Stream flow and concentration data, including stream flow discharge, temperature, specific conductance, turbidity, dissolved oxygen, nitrate concentration, and phosphorus concentration, is collected by the United States Geologic Survey (USGS). The gauge number and the corresponding watersheds for different variables can be found in the Supporting Information S1. The data can be accessed from CUAHSI HydroShare (http://doi.org/10.4211/hs.a70e86faebe448128ecb0e208626ba4d). The codes used for compressed sensing can be found from the "Data Driven Science & Engineering—Machine Learning, Dynamical Systems, and Control" book by Brunton and Kutz (2022) (http://www.databookuw.com/).

**Acknowledgments**
This work relates to Department of Army award ("Novel Technologies to Mitigate Water Contamination for Resilient Infrastructure"; Federal Award Identification Number: W9132T2220001) issued by the U. S. Army Corps of Engineers (USACE) Engineer Research and Development Center (ERDC). The United States Government has a royalty-free license throughout the world in all copyrightable material contained herein. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of USACE-ERDC.

## References

Banner, E. B. K., Stahl, A. J., & Dodds, W. K. (2009). Stream discharge and riparian land use influence In-stream concentrations and loads of phosphorus from central plains watersheds. *Environmental Management*, *44*(3), 552–565. https://doi.org/10.1007/S00267-009-9332-6/TABLES/6

Bremer, L. L., Hamel, P., Ponette-González, A. G., Pompeu, P. V., Saad, S. I., & Brauman, K. A. (2020). Who are we measuring and modeling for? Supporting multilevel decision-making in watershed management. *Water Resources Research*, *56*(1), e2019WR026011. https://doi.org/10.1029/2019WR026011

Brunton, S., & Kutz, J. (2022). Data-driven science and engineering: Machine learning, dynamical systems, and control. https://doi.org/10.1017/9781108380690.002

Burns, D. A., Pellerin, B. A., Miller, M. P., Capel, P. D., Tesoriero, A. J., & Duncan, J. M. (2019). Monitoring the riverine pulse: Applying high-frequency nitrate data to advance integrative understanding of biogeochemical and hydrological processes. *Wiley Interdisciplinary Reviews: Water*, *6*(4), 1–24. https://doi.org/10.1002/wat2.1348

Candès, E. J., Romberg, J., & Tao, T. (2006). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, *52*(2), 489–509. https://doi.org/10.1109/TIT.2005.862083

Candes, E. J., & Wakin, M. B. (2008). An introduction to compressive sampling. *IEEE Signal Processing Magazine*, *25*(2), 21–30. https://doi.org/10.1109/msp.2007.914731

Cassidy, R., & Jordan, P. (2011). Limitations of instantaneous water quality sampling in surface-water catchments: Comparison with near-continuous phosphorus time-series data. *Journal of Hydrology*, *405*(1–2), 182–193. https://doi.org/10.1016/j.jhydrol.2011.05.020

Chatterjee, A. (2000). An introduction to the proper orthogonal decomposition. *Current Science*, 808–817. https://www.jstor.org/stable/24103957%23metadata%5Finfo%5Ftab%5Fcontents

Crawford, J. T., Loken, L. C., Casson, N. J., Smith, C., Stone, A. G., & Winslow, L. A. (2015). High-speed limnology: Using advanced sensors to investigate spatial variability in biogeochemistry and hydrology. *Environmental Science and Technology*, *49*(1), 442–450. https://doi.org/10.1021/es504773x

Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, *52*(4), 1289–1306. https://doi.org/10.1109/TIT.2006.871582

Gammons, C. H., Babcock, J. N., Parker, S. R., & Poulson, S. R. (2011). Diel cycling and stable isotopes of dissolved oxygen, dissolved inorganic carbon, and nitrogenous species in a stream receiving treated municipal sewage. *Chemical Geology*, *283*(1–2), 44–55. https://doi.org/10.1016/J.CHEMGEO.2010.07.006

Gao, L., Qi, J., Li, S., Benoy, G., Xing, Z., & Meng, F. R. (2020). Effects of sampling frequency on estimation accuracies of annual loadings for water quality parameters in different sized watersheds. *Water Quality Research Journal*, *55*(3), 261–277. https://doi.org/10.2166/wqrj.2020.012

Jones, A. S., Horsburgh, J. S., Mesner, N. O., Ryel, R. J., & Stevens, D. K. (2012). Influence of sampling frequency on estimation of annual total phosphorus and total suspended solids loads. *Journal of the American Water Resources Association*, *48*(6), 1258–1275. https://doi.org/10.1111/j.1752-1688.2012.00684.x

Katul, G. G., Porporato, A., Daly, E., Oishi, A. C., Kim, H. S., Stoy, P. C., et al. (2007). On the spectrum of soil moisture from hourly to interannual scales. *Water Resources Research*, *43*(5), 1–10. https://doi.org/10.1029/2006WR005356

Kerr, J. G., Zettel, J. P., McClain, C. N., & Kruk, M. K. (2018). Monitoring heavy metal concentrations in turbid rivers: Can fixed frequency sampling regimes accurately determine criteria exceedance frequencies, distribution statistics and temporal trends? *Ecological Indicators*, *93*, 447–457. https://doi.org/10.1016/j.ecolind.2018.05.028

Kirchner, J. W., & Neal, C. (2013). Universal fractal scaling in stream chemistry and its implications for solute transport and water quality trend detection. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(30), 12213–12218. https://doi.org/10.1073/pnas.1304328110

Lustig, M., & Donoho, D. (2008). Compressed sensing MRI. In *Signal processing magazine* (pp. 72–82).

Manohar, K., Brunton, B. W., Kutz, J. N., & Brunton, S. L. (2018). Data-driven sparse sensor placement for reconstruction: Demonstrating the benefits of exploiting known patterns. *IEEE Control Systems*, *38*(3), 63–86. https://doi.org/10.1109/MCS.2018.2810460

Minaudo, C., Dupas, R., Gascuel-Odoux, C., Fovet, O., Mellander, P. E., Jordan, P., et al. (2017). Nonlinear empirical modeling to estimate phosphorus exports using continuous records of turbidity and discharge. *Water Resources Research*, *53*(9), 7590–7606. https://doi.org/10.1002/2017WR020590

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Needell, D., & Tropp, J. A. (2009). CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, *26*(3), 301–321. https://doi.org/10.1016/j.acha.2008.07.002

Ohmer, M., Liesch, T., & Wunsch, A. (2022). Spatiotemporal optimization of groundwater monitoring networks using data-driven sparse sensing methods. *Hydrology and Earth System Sciences*, *26*(15), 4033–4053. https://doi.org/10.5194/HESS-26-4033-2022

Parolari, A. J., Sizemore, J., & Katul, G. G. (2021). Multiscale legacy responses of soil gas concentrations to soil moisture and temperature fluctuations. *Journal of Geophysical Research: Biogeosciences*, *126*(2), 1–16. https://doi.org/10.1029/2020JG005865

Pellerin, B. A., Bergamaschi, B. A., Gilliom, R. J., Crawford, C. G., Saraceno, J., Frederick, C. P., et al. (2014). Mississippi river nitrate loads from high frequency sensor measurements and regression-based load estimation. *Environmental Science and Technology*, *48*(21), 12612–12619. https://doi.org/10.1021/es504029c

Pluth, T. B., Brose, D. A., Gallagher, D. W., & Wasik, J. (2021). Long-term trends show improvements in water quality in the chicago metropolitan region with investment in wastewater infrastructure, deep tunnels, and reservoirs. *Water Resources Research*, *57*(6), 1–20. https://doi.org/10.1029/2020WR028422

Reynolds, K. N., Loecke, T. D., Burgin, A. J., Davis, C. A., Riveros-Iregui, D., Thomas, S. A., et al. (2016). Optimizing sampling strategies for riverine nitrate using high-frequency data in agricultural watersheds. *Environmental Science and Technology*, *50*(12), 6406–6414. https://doi.org/10.1021/acs.est.5b05423

Rode, M., Wade, A. J., Cohen, M. J., Hensley, R. T., Bowes, M. J., Kirchner, J. W., et al. (2016). Sensors in the stream: The high-frequency wave of the present. *Environmental Science and Technology*, *50*(19), 10297–10307. https://doi.org/10.1021/acs.est.6b02155

Rozemeijer, J. C., Van Der Velde, Y., Van Geer, F. C., De Rooij, G. H., Torfs, P. J. J. F., & Broers, H. P. (2010). Improving load estimates for $NO_3$ and P in surface waters by characterizing the concentration response to rainfall events. *Environmental Science and Technology*, *44*(16), 6305–6312. https://doi.org/10.1021/es101252e

Skeffington, R. A., Halliday, S. J., Wade, A. J., Bowes, M. J., & Loewenthal, M. (2015). Using high-frequency water quality data to assess sampling strategies for the EU Water Framework Directive. *Hydrology and Earth System Sciences*, *19*(5), 2491–2504. https://doi.org/10.5194/hess-19-2491-2015

Smith, R. A., Schwarz, G. E., & Alexander, R. B. (1997). Regional interpretation of water-quality monitoring data. *Water Resources Research*, *33*(12), 2781–2798. https://doi.org/10.1029/97WR02171

Thompson, J., Pelc, C. E., & Jordan, T. E. (2021). Water quality sampling methods may bias evaluations of watershed management practices. *Science of the Total Environment*, *765*, 142739. https://doi.org/10.1016/j.scitotenv.2020.142739

Vasseur, D. A., & Yodzis, P. (2004). The color of environmental noise. *Ecology*, *85*(4), 1146–1152. https://doi.org/10.1890/02-3122

Wei, J., Wang, L., Liu, P., Chen, X., Li, W., & Zomaya, A. Y. (2017). Spatiotemporal fusion of MODIS and landsat-7 reflectance images via compressed sensing. *IEEE Transactions on Geoscience and Remote Sensing*, *55*(12), 7126–7139. https://doi.org/10.1109/TGRS.2017.2742529

Zabaleta, A., Martínez, M., Uriarte, J. A., & Antigüedad, I. (2007). Factors controlling suspended sediment yield during runoff events in small headwater catchments of the Basque Country. *Catena*, *71*(1), 179–190. https://doi.org/10.1016/J.CATENA.2006.06.007

Zhang, L., Xing, M., Qiu, C. W., Li, J., Sheng, J., Li, Y., & Bao, Z. (2010). Resolution enhancement for inversed synthetic aperture radar imaging under low SNR via improved compressive sensing. *IEEE Transactions on Geoscience and Remote Sensing*, *48*(10), 3824–3838. https://doi.org/10.1109/TGRS.2010.2048575