An Assessment of Annual Load Estimation Methods in Small Watersheds for Cross Site Comparisons

***Nicholas J Gubbins*[1*]**, *Weston M Slaughter*[2], *Michael J Vlah*[3], *Spencer Rhea*[3], *William H McDowell*[4,5], *Emily S Bernhardt*[3], *Matthew RV Ross*[1]

(1) Department of Ecosystem Science and Sustainability, Colorado State University, Fort Collins, CO; (2) Department of Geology, University of Maryland, College Park, MD; (3) Department of Biology, Duke University, Durham, NC; (4) Department of Natural Resources and the Environment, University of New Hampshire Main Campus, Durham NH; (5) Institute of Environment, Florida International University, Miam FL

*Corresponding author: contact gubbins@colostate.edu

Abstract

Rivers and streams transport dissolved materials and sediments from their watersheds, providing researchers and land managers with insights into water quality, watershed landscape health, and management strategies. Despite the widespread use of export (load) estimates, comparisons at macroscales are complicated by differences in underlying data quality and estimation methods between sites, periods, and solutes. Using high-frequency data from the Hubbard Brook Experimental Forest and the Plynlimon Research Catchments, we generated time series of increasingly coarse sampling frequencies, and tested the sensitivity of various load estimation methods. We further tested the accuracy of common methods using synthetic time series, spanning a range of flow regimes and concentration-discharge (C:Q) relationships. Lastly, we applied each estimation method to the MacroSheds dataset (macrosheds.org), generating a publicly available dataset of 16,489 site-years across 93 sites and 112 solutes. Results indicate that load estimates with high sampling frequency (daily or better) and an informative concentration-discharge relationship are well suited for macroscale science efforts (errors within ~10%). Estimates based on coarse (biweekly or coarser) underlying data and incompletely described and/or complex C:Q relationships showed large enough error (>50%) to suggest they would be misleading if included in macroscale efforts. Our results suggest that scientists interested in comparing load estimates should first consider (1) sensitivity of their analysis to changes in load magnitudes, (2) the underlying data frequency used to generate estimates, (3) the C:Q relationship of their solute of interest, and (4) their confidence in the completeness of that C:Q relationship over the period of study.

Data Availability

Water chemistry and streamflow data for this paper were sourced from MacroSheds (Vlah et al, 2023, macrosheds.org), Hubbard Brook Experimental Forest (Hubbard Brook Watershed Ecosystem Record, 2023), and Plynlimon Research Catchments (Colin et al., 2012). Versioned code and the underlying dataset and workflow used for the analysis is available at https://github.com/ecogub/RSFME/releases/tag/v0. Resulting load calculations are hosted publicly at 10.6084/m9.figshare.24975504.

1 Introduction

Annual stream load is the mass of solutes or sediments that moves past a point in a stream during a water year. Quantifying the magnitude, timing and form of solute exports from watershed ecosystems gives researchers key insights into how a watershed functions as a system –such as a watershed's capacity to retain nutrients and potential to weather bedrock. In small watersheds with relatively watertight bedrock, researchers can assume that water only leaves the watershed through evaporation, transpiration, or stream discharge (Bormann et al., 1968). If this assumption is met, researchers interested in how nutrients move through ecosystems can measure nutrient inputs from the atmosphere and soil and then measure the outputs of these nutrients in streamflow. The difference between inputs to and outputs from a small watershed can be used to estimate critical ecosystem functions, such as nutrient retention, precisely and accurately (Bormann et al., 1968). However, this kind of chemical accounting relies on accurate estimates of inputs and outputs. For instance, researchers interested in nutrient dynamics rely on estimates of solute export to compare watershed functions in paired catchment studies (such as Bormann et al., 1968; Likens et al., 1970; Likens et al., 2006). Geochemists use load estimates to constrain in-watershed weathering rates (Gaillardet et al, 1999; Maher and Chamberlain, 2014). Land managers and government decision makers rely on – and invest in – accurate estimation of solute loads to write and apply policy, such as fishery management, water treatment, and conservation efforts (Meals et al., 2013; Dodds et al., 2008; Schilling et al., 2017).

Ideally, solute loads would be calculated as the product of the solute's concentration and streamflow, integrated *continuously* over time, as shown in Equation 1, where L is the load, C the concentration, Q the streamflow, and t the time.

$$(Equation\ 1)\quad L = \int C(t)Q(t)dt$$

In reality, environmental data are rarely truly continuous. However, after a century of collective effort, streamflow gauging and modelling can generate reliable, near-continuous estimates of discharge for most systems of interest. To measure flux at this frequency, however, would require solute concentration measurements to be near-continuous as well. Yet this kind of data, whether directly measured or modelled, is rare due to the labor and time involved in producing the measurements (Kirchner et al., 2004; Pellerin et al., 2014; Zimmer et al., 2019). Until recently (Halliday et al., 2012), measuring most solutes required a water sample to be taken to a laboratory. As a result, chemistry samples have traditionally been taken as discrete “grab” samples that must be manually collected and analyzed off-site (Kirchner et al., 2004; Pellerin et al., 2014). Sampling frequencies, therefore, have been on the order of weekly or monthly for many watershed-ecosystem studies for more than 50 years (Buso et al., 2000).

Estimating solute loads for studies with infrequent chemical sampling requires the use of models to estimate concentrations between measurements (Kirchner et al., 2004; Pellerin et al., 2014). This introduces uncertainty into truly continuous estimates of load (Richards and Holloway, 1987; Schilling et al., 2017). In recent decades, the technology to monitor water chemistry using high-frequency environmental sensors has become cost-effective enough –for a limited range of solutes and solute proxy measurements– to be measured at the same frequencies at which discharge is modeled (Kirchner et al., 2004; Pellerin et al., 2014). However, for researchers interested in historical solute records or solutes that can’t yet be measured using sensors, there is no single, ideal method for assessing the accuracy and potential bias of load estimates (Richards and Holloway, 1986; Appling et al., 2015; Aulenbach et al., 2016). Moreover, many different methods have been used to compute these estimates (Figure 1), ranging from simple averaging or step functions (Likens et al., 1977) to complex, data-intensive statistical models (Appling et al., 2015; Zhang and Hirsch, 2019). Conflicting estimation methods further challenge confidence in comparing load estimates across sites (Appling et al., 2015; Nava et al., 2019).

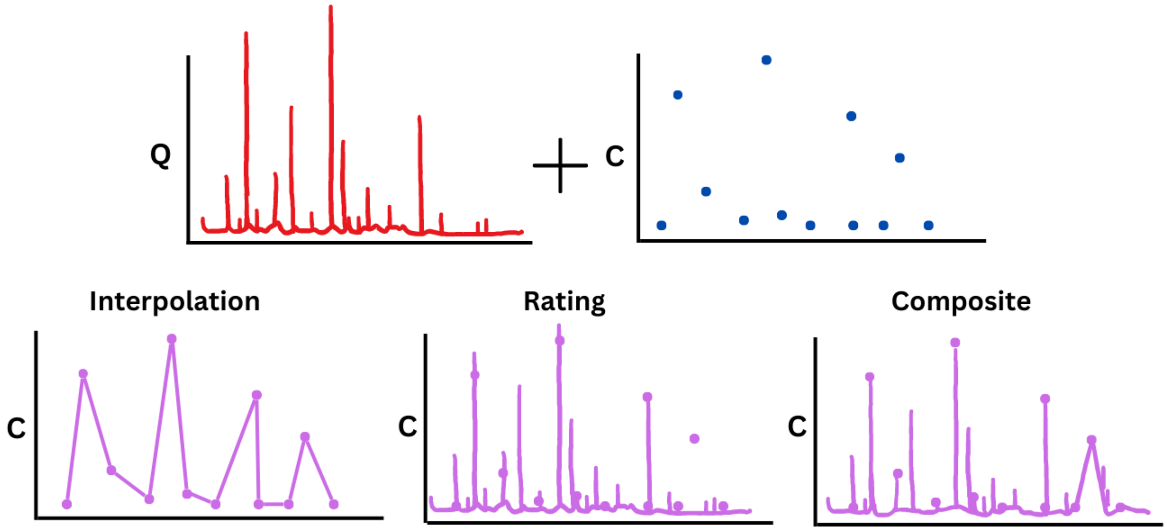

Figure 1: An illustration of concentration estimation techniques (in purple) resulting from high-frequency discharge data (in red) and intermittent, discrete chemistry samples (in blue). Interpolation assumes missing concentrations are predicted best by linearly interpolating observed concentrations. The rating method assumes a strong relationship between concentration and discharge, then uses discharge to impute missing concentrations. The composite method

uses the rating method, but then “forces” the imputed concentrations through all known observations. The Beale ratio estimator (not shown) instead applies a correction factor after generating an estimate with available data.

While some methods are more appropriate than others for a given system and solute, it is impossible to know with confidence what was missed in years with infrequent sampling (Richards and Holloway, 1986; Appling et al., 2015; Nava et al., 2019). This puts water researchers, especially those interested in macroscale, cross-site comparisons, in a quandary, leading to the core question of this paper:

How do we make accurate and analysis-ready load estimates that can be compared across diverse watersheds, sampling regimes, and data density?

The challenges outlined above are exacerbated when expanding scope from a single site and solute to many of each, where methodologies and sampling frequency can vary substantially. Thus, answering our primary research question becomes especially important for macrosystem scientists (researchers interested in working across many systems), as they are often unable to rely on the intimate, site-specific knowledge that can inform single-site analyses. While the fundamental problems presented here cannot necessarily be solved, many recent efforts have harmonized previously scattered watershed data. For example, the CAMELS dataset is a compilation of high quality hydrology, geospatial, and climate data for well-sampled USGS sites (Newman et al., 2014). The work the CAMELS team has done to harmonize and document these data has significantly lightened the repetitive, pre-analysis burden that macrosystem scientists face.

Many efforts have been taken to improve the accuracy of solute load estimates and provide recommendations for approaches (Richards and Holloway, 1986; Appling et al. 2015; Aulenbach et al., 2016; Shilling et al., 2017; etc.), but those efforts have either focused on large watersheds (Pellerin et al., 2014, Appling et al., 2016; Schilling et al., 2017) or on choosing the best method possible for a given time series (Richards and Holloway, 1986; Aulenbach et al., 2016). While both areas of research are deeply important for the field at large, they do not clearly delineate when load estimates can be used in cross-site comparisons. Additionally, robust statistical methods have been developed, such as the USGS’s Weighted Regressions on Time, Discharge, and Season (WRTDS) tool (Hirsch et al., 2010) that accurately model loads at sites with at least decadal records (Hirsch et al., 2010). However, these methods require large amounts of data and site specific tuning and model generation that may be prohibitively time consuming across large, spatially diverse datasets. This study seeks to build on these prior efforts with three goals:

1. To provide clear guidance to macroscale watershed scientists about when load estimates from small studies are interoperable.

2. To provide flux estimates for a synthesis dataset of small watershed studies.

3. To provide a framework for classifying potentially comparable flux estimates for cross-site analyses.

To reach our goals, we focus on three sub-questions:

1. How sensitive are commonly used load estimation methods to varying frequencies of concentration sampling?

2. How does this sensitivity scale with the variance and drivers of the concentrations sampled?

3. Can we reliably choose the best available estimation method using only solute concentration and discharge data?

2 Methods

To tackle the first sub-question, our study used data from Hubbard Brook Experimental Forest (HBEF) in New Hampshire, USA (Hubbard Brook Watershed Ecosystem Record, 2023) and Plynlimon Research Catchments in Wales, UK (Colin et al., 2012) to show the relative effect of estimation method and sampling frequency on load estimation error for two solutes. We then applied our case study methods on synthetic time series data to explore the second sub-question across a range of hydrologic regimes and solute chemodyamics. To explore sub-question three, we tested a simplified application of a published decision tree for method selection. Lastly, we applied our tested methods to the geographically diverse MacroSheds synthesis dataset (Vlah et al., 2023), providing a dataset of load estimates to be used for future investigations.

2.1 Load estimation

Four common load methods were chosen for this study: linear interpolation (LI), Beale ratio, rating, and composite. These methods are archetypal of the array methods commonly used in small-watershed ecosystem studies. Previous work has shown that methods should be matched to the chemodynamics and data density of the time series of interest (Aulenbach et al., 2016).

2.1.1 Linear interpolation

In load estimation by linear interpolation (Figure 1, bottom left), sampled chemistry values are interpolated to match the sampling frequency of their accompanying discharge time series. Load is then computed according to Equation 2.

$$(Equation\ 2) \quad L = \sum_{i=1}^{n} C_i^{int} Q_i$$

Where L is the load in kg/ha/year, $C_i^{int}$is the interpolated concentration in mg/L, and is Q the streamflow in L/s. Due to its simplicity, linear interpolation is commonly used in studies, especially where load estimation methods are not the focus. The method has been shown to work well for frequently sampled time series (Appling et al., 2015) and in time series where concentrations are highly autocorrelated (Aulenbauch et al., 2016). For example, this approach has been used to estimate fluxes of nutrients and suspended sediments in the Santa Ynez Mountains (Coombs and Melack, 2012) and in describing solute export from catchments in the Bonanza Creek LTER study site (Petrone et al., 2006). We used the “RiverLoad” R package’s “method6” function to apply linear interpolation (Nava et al., 2019).

2.1.2 Beale ratio estimator

The Beale ratio estimator is often chosen for its ability to provide unbiased load estimates (Meals et al., 2013; Nava et al., 2019). This method uses covariance of load and discharge to scale load estimates for a given period as represented in Equation 3.

$$(equation\ 3)\quad L = Q\frac{\underline{l}}{\underline{Q}}\left[\frac{1 + n\left[\frac{Cov(l,Q)}{\underline{l}\,\underline{Q}}\right]}{1 + n\left[\frac{Cov(l,Q)}{\underline{Q}^2}\right]}\right]$$

Where Q is the flow, l the instantaneous load, $\underline{l}$ is the total load for when concentration was measured, and $\underline{Q}$ is the total of flow for the entire year. We used the “RiverLoad” R package’s “beale.ratio” function to apply the Beale ratio estimator, which relies on the Beale ratio as described in Beale (1962). This approach is often used by the US Environmental Protection Agency (EPA) in estimating loads (Meals et al., 2013). Current use of the Beale ratio estimator is often improved by the EPA’s AutoBeale tool, which employs an algorithm to detect the most appropriate time windows over which to calculate Beale ratios in a given dataset (Lee et al., 2019).

2.1.3 Rating

The rating method (Figure 1, bottom middle) first relates concentration to discharge in log-log space with a simple linear model. Then the resulting least-squares regression line is used to generate a full time series of concentrations using the discharge time series as an input. Values are then summed for the year to generate load, just as in linear interpolation. This method has been shown to be very effective when the solute of interest has a strong C:Q relationship (Crawford, 1996; Quilbe et al., 2006). We used the “RiverLoad” R package”s “rating” function to generate rating estimates. This method has been used to estimate nutrient loading in the Illinois River (Vieux and Moreda, 2003) and export in urban streams in Puerto Rico (McDowell et al., 2019). In cases where the rating method is appropriate, current efforts generally favor more sophisticated regression methods, such as the WRTDS tool (Lee et al., 2019).

2.1.4 Composite

The composite method (Figure 1, bottom right) follows the same steps as the rating method, but the resulting daily concentration time series is corrected back to the observed values used to generate the rating. First, a rating is fit as described in the previous section. Then, residuals between each observation and the rating-derived concentration at that time are calculated. The residuals are then applied as a correction to the derived time series at each sampled point. Each residual correction is then linearly interpolated between sampled days to generate a final time series. In essence, the rating-generated time series is “forced through” all known observations. This time series is then summed to compute annual loads. The composite method has been shown to combine the strengths of linear interpolation and the rating method (Aulenbach and Hooper, 2006; Appling et al., 2015; Aulenbach et al., 2016). Our application of the composite method was adapted from Appling et al. 2015. The composite method has become a premier choice for many loading analyses. For example, it has been used to estimate sulfate export at Niwot Ridge LTER site in Colorado (Crawford, 2019) and is often favored by the US Geological Survey (Zhang and Hirsch, 2019).

2.2 Comparing subsampled data to high-frequency data

To estimate each method’s sensitivity to solute and sampling frequency, we performed a data coarsening experiment on four time series. Nitrate as nitrogen (NO_{3}-N) and calcium (Ca, Figure 2) from the Watershed 3 site at HBEF (40.2 hectare watershed area) and the Upper Hafren site (120 ha watershed area) at the Plynlimon Research catchments.

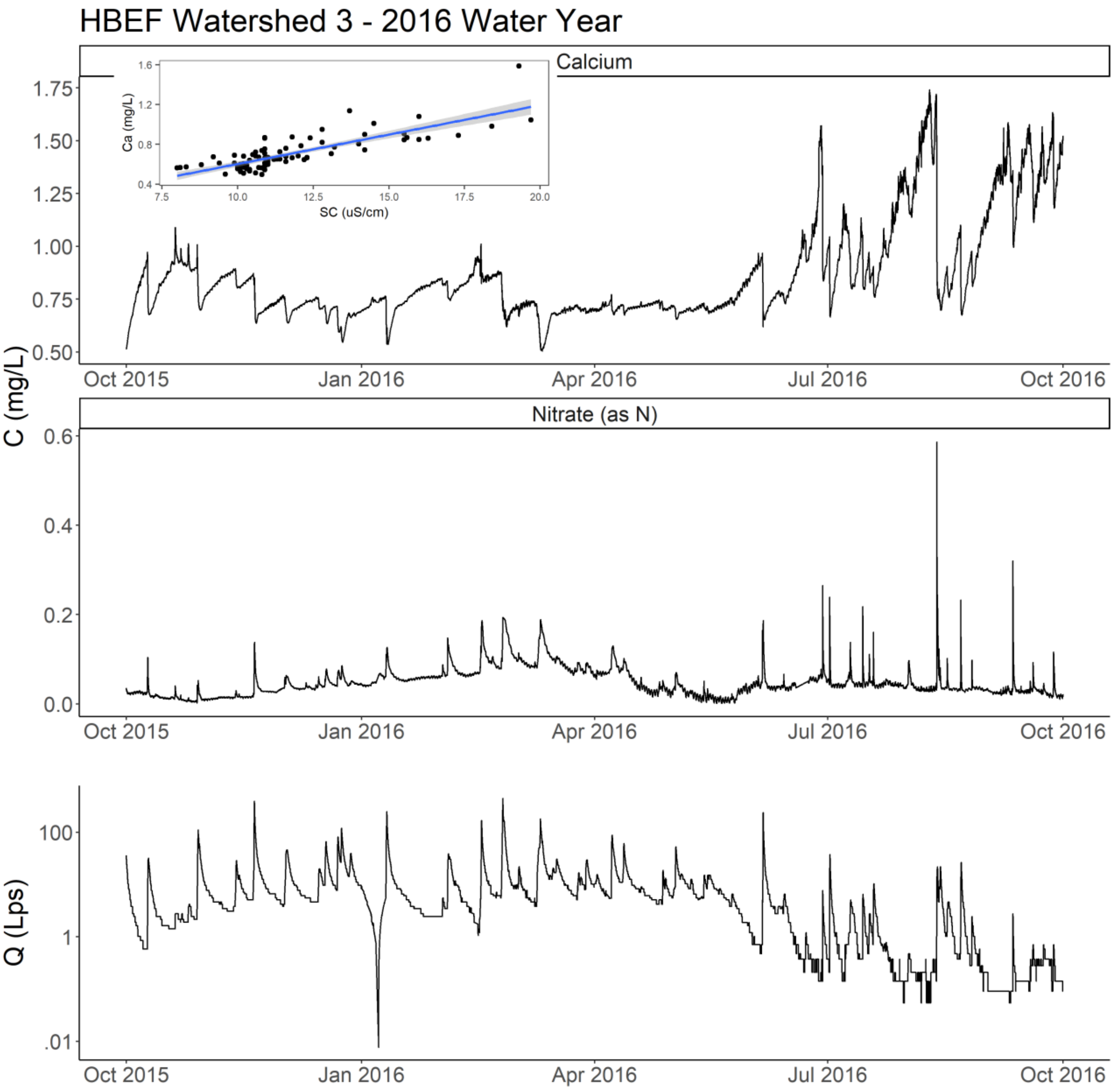

Figure 2: The underlying chemistry and streamflow time series used in the data coarsening experiments. Chemistry data was collected at Hubbard Brook Experimental Forest in watershed 3 using a multiparameter sonde and UV-VIS nitrate analyzer. The NO_3-N time series was measured by the nitrate analyzer directly. The Ca time series was derived from sonde measurements of specific conductance regressed against grab samples of calcium. The specific conductance record has no missing days. Streamflow was collected using a long-running rating, v-notch weir, and stage recorder.

Both HBEF chemistry time series were collected over the 2016 water year at a 15-minute frequency using a multiparameter sonde in conjunction with a long-running stream gauge. The Ca time series was constructed by fitting a least square regression line with no intercept between specific conductance sensor readings and Ca grab samples taken at the site (Figure 2, inset). The NO_3-N time series had a mean concentration of 0.048 mg/L with a standard deviation of 0.032 mg/L. The Ca time series had a mean concentration of 0.86 mg/L with a standard deviation of 0.23 mg/L. The Ca time series was complete for the year, with no missing days in the underlying specific conductance dataset. The NO_3-N time series had 3 days with incomplete data (2/25/2016, 6/18/2016, and 6/23/2016) and 4 days with no data (6/19/2016-6/22/2016).

The HBEF discharge time series (figure 2) is complete, with a mean flow of 8.94 Lps, a flow standard deviation of 19.20 Lps, and a yield of 2.6 x 10^{10} liters for the year. The steep dip between 1/2/2016 and 1/10/2016 was caused by freezing at the site. Hydrologists at HBEF use the robust ensemble method for imputing streamflow during freezes by using data from nearby, non-frozen gauges, as described in See et al. (2020).

The underlying C:Q relationships for both solutes over the year at HBEF are presented in figure 3.

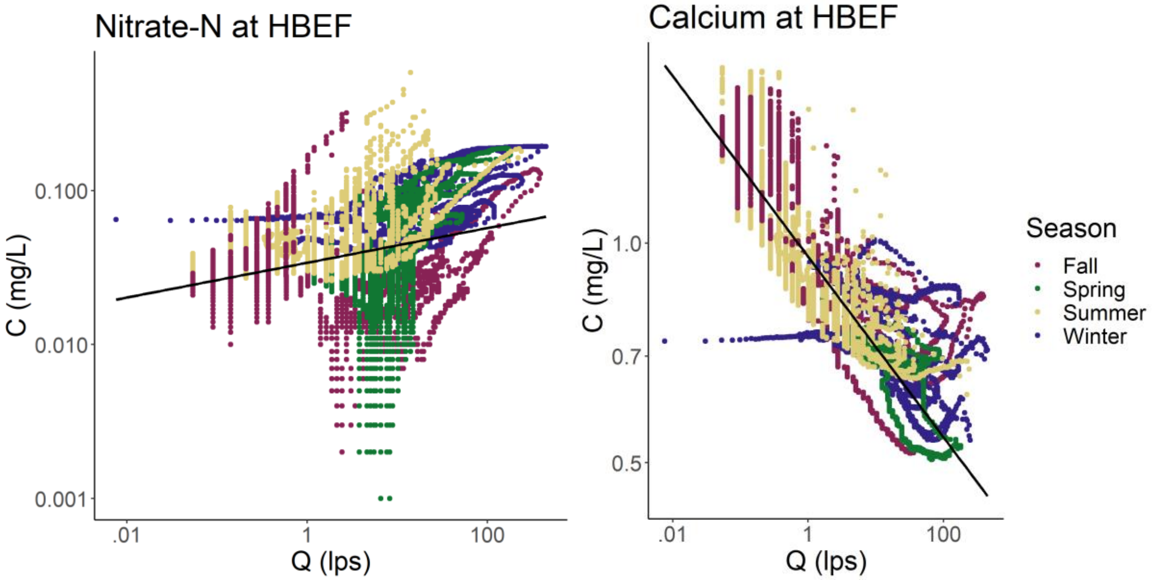

Figure 3: Concentration-discharge plot for NO_3-N and Ca at HBEF watershed 3 for the 2016 water year. Observations are colored by season and the least-squares regression line used in the rating and composite methods is in black. Note the presence of hysteresis loops and seasonality that our methods do not handle. The slope of the best fit line is 0.11 with an r-squared of 0.07 for NO_3-N and -0.12 with an r-squared of 0.79 for Ca. The NO_3-N time series has a mean of 0.048 mg/L and a standard deviation of 0.032 mg/L. The Ca time series has a mean of 0.86 mg/L and a standard deviation of 0.22 mg/L.

There is a very weak relationship between NO_3-N and discharge at watershed 3, with a log-log least-squares regression slope of 0.11 and r-squared of 0.07. In contrast, Ca showed a strong, diluting C:Q relationship, with a log-log least-squares regression slope of -0.12 and r-squared of 0.79.

Both chemistry time series from Plynlimon were collected every seven hours over the 2008 water year using auto-samplers and a long running stream gauge (Figure 4) (Kirchner and Reynolds, 2013). The Ca time series had a mean of 0.52 mg/L and a standard deviation of 0.12 mg/L. The NO3-N time series had a mean of 0.14 mg/L and a standard deviation of 0.06 mg/L. Data at Plynlimon was both measured more sparsely and had more gaps than the HBEF data. The Ca time series had 20 incomplete days and the nitrate time series has 37 incomplete days.

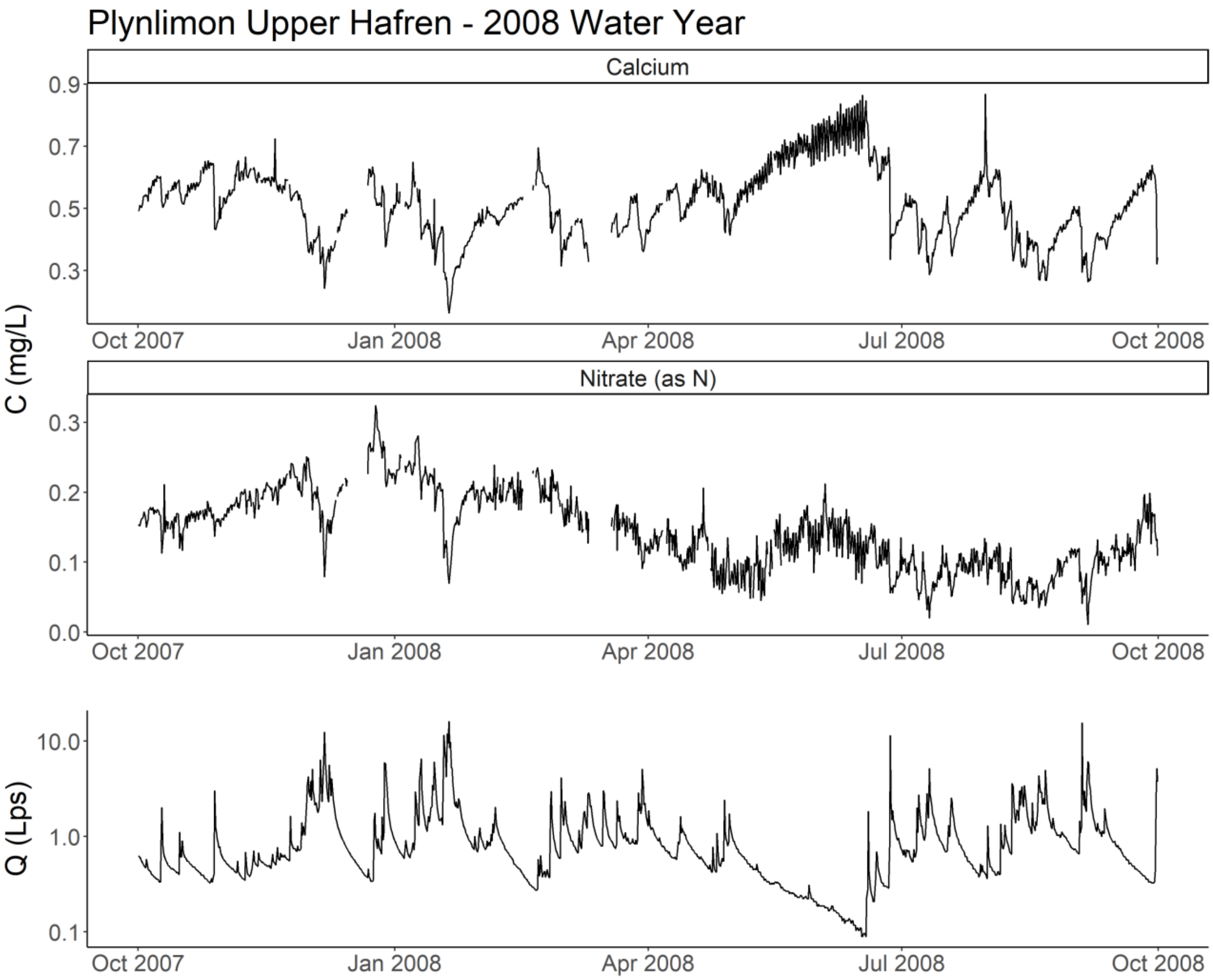

Figure 4: A chemistry time series collected at Plynlimon's Upper Hafren site using grab samples. Note the many small gaps in both time series. Streamflow was collected using a long-running rating, a weir, and a stage recorder. The flow record is complete for the year.

The underlying streamflow record at the Upper Hafren site (figure 4) was complete, with no missing days. The flow record had a mean of 1.08 Lps, a standard deviation of 1.36 Lps, and a yield of 3.4 x 10^7 liters for the year.

C:Q relationships at Plynlimon generally mirrored those at Hubbard Brook. Namely, NO_3-N shows a complex, seasonal C:Q relationship with a weak fit and Ca shows a more consistent, diluting relationship with a stronger fit. However, both fits are less explanatory at Plynlimon than at Hubbard Brook. NO_3-N also shows a slight diluting trend at Plylimon (best fit slope of -0.01) compared to the slight enriching trend at Hubbard Brook (best fit slope of 0.11). Both C:Q relationships are presented in figure 5.

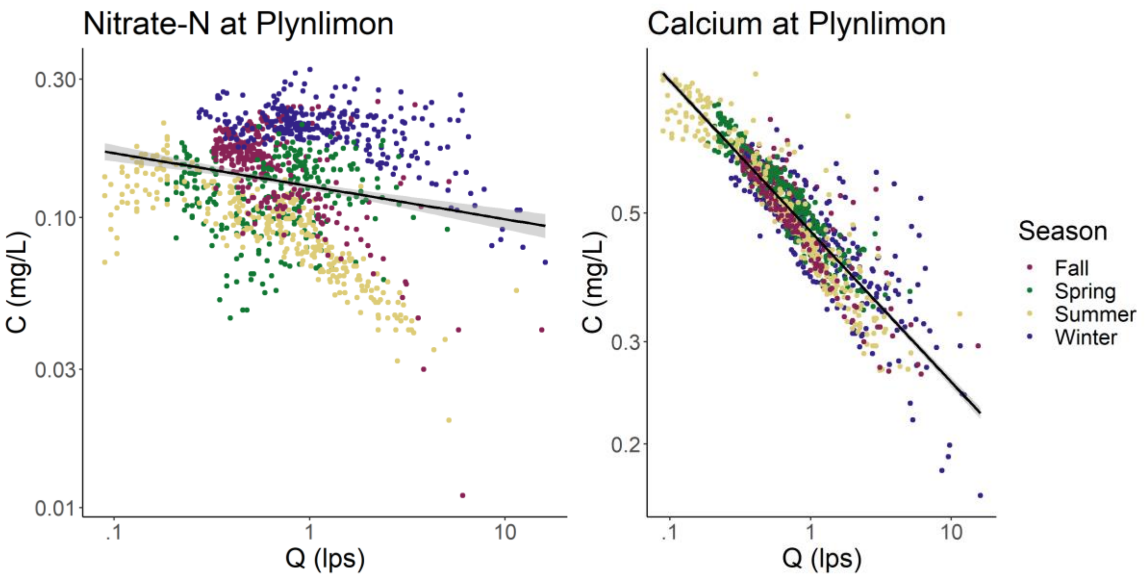

Figure 5: Concentration-discharge plot for NO_3-N and Ca at Plynlimon's Upper Hafren site for the 2008 water year. Observations are colored by season and the least-squares regression line used in the rating and composite methods is in black. The slope of the best fit line is -0.01 with an r-squared of 0.05 for NO3-N and -0.05 with an r-squared of 0.40 for Ca. The NO_3-N time series has a mean of 0.14 mg/L and a standard deviation of 0.06 mg/L. The Ca time series has a mean of 0.52 mg/L and a standard deviation of 0.12 mg/L.

2.2.1 Coarsening procedure

For Hubbard Brook, we coarsened each time series from the full, 15-minute resolution to daily by hour (sampled hourly, every other hour, every third hour, …, every 24th hour), from daily to weekly by day (sampled every day, every other day, every third day, …, every seventh day), and then to monthly and bimonthly discretely, as partially illustrated in Figure 6. A similar methodology was applied to the Plynlimon time series data, excluding coarsening intervals finer than the 7-hour sampling frequency.

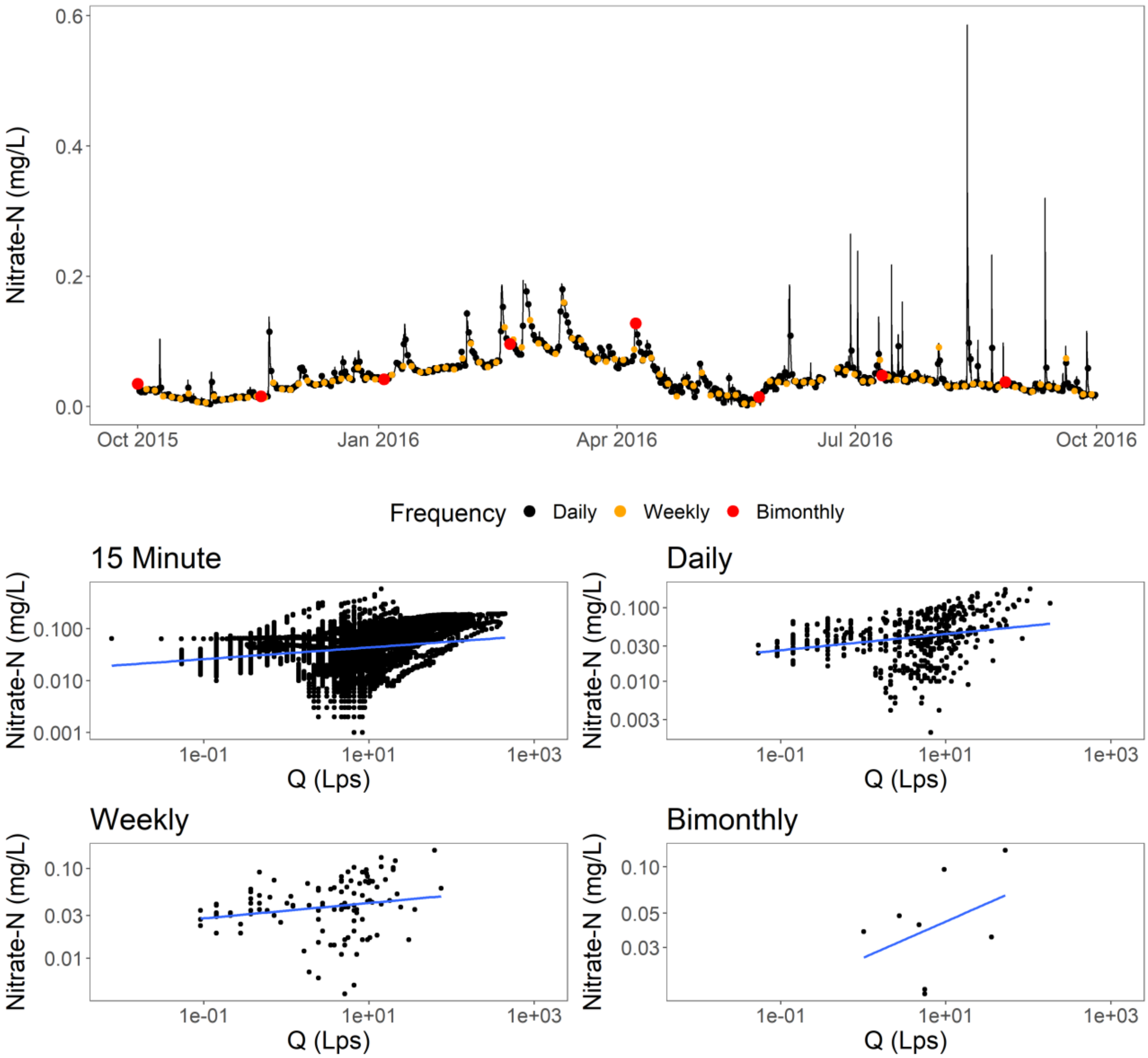

Figure 6: A selection of gradually coarsened nitrate time series generated from the original, 15-minute HBEF sensor time series (black line). Coarsened daily, weekly, and bimonthly are shown as an example. The resulting concentration-discharge relationships are shown below. This process was repeated with a random starting position 100 times to create a range of possible time series from the original data.

A random starting point from within each initial coarsening interval was chosen and then every nth sample was then taken to create a coarsened time series. We then applied each method to each coarsened time series and the full discharge record to generate annual flux estimates. This process was repeated 100 times to generate an envelope of possible estimates with each method.

2.2.2 Calculating and comparing to “true” load

“True” load was calculated by applying the composite method to the full, high frequency time series (as recommended in Aulenbach et al. 2016). We then calculated percent error by comparing true load and estimates generated with the coarsened time series.

2.3 Comparing model efficacy for synthetic datasets with varying C:Q relationships

Due to the limited availability of high-quality, high-frequency chemistry and streamflow data from small watersheds, we created synthetic time series to test the sensitivities of each method to various hydrologic and C:Q relationships.

Past work has shown that the best available load estimation method depends largely on the autocorrelation of chemistry samples through time– driven either by sampling frequency or in situ processes–and the relationship between solute concentration and discharge (Aulenbach et al., 2016). The variance (including diurnal, seasonal, and interannual) and C:Q relationship of a given solute over time define its chemodynamics. Solutes with high variance and/or with strong C:Q relationships are considered chemodynamic (Godsey et al., 2009; Koger et al., 2018; Godsey et al., 2019). Stable, discharge-invariant solutes are considered chemostatic (Godsey et al., 2009; Koger et al., 2018; Godsey et al., 2019). Chemodynamism has been observed in many forms. Some commonly chemodynamic solutes, like dissolved organic matter, often increase with discharge, and are termed “enriching” (Moatar et al., 2017). The opposite is true of many geochemical solutes, such as magnesium or potassium. Instead, they often dilute as flows increase (Moatar et al., 2017; Godsey et al., 2009). Additionally, solutes can display complex chemodynamics that change as flows increase (Moatar et al., 2017). Solutes can also display no pattern with streamflow, but vary widely due to other factors, such as environmental biotic demand. This is often true of nutrients, such as nitrate (Pellerin et al., 2014; Schilling et al., 2017). Given the underlying assumptions used in the estimation methods outlined above, understanding the underlying chemodynamics of a given solute is critical to selecting an appropriate method (Appling et al., 2015; Aulenbach et al., 2016).

Using the discharge record from HBEF Watershed 3 in the 2016 water year as a starting point, we created batches of three idealized hydrologic regimes and four idealized C:Q relationships to test our flux methods. We have grounded our analysis in known methods where possible and the code used to create the data is publicly available at https://github.com/ecogub/RSFME/releases/tag/v0.

2.3.1 Hydrologic regimes

We fit an autoregressive integrated moving average (ARIMA) model to the flow record used in the data coarsening experiments using the “forecast” R package’s “auto.arima” function (Khandakar, 2008). Then, residuals from the model were resampled (with replacement) and applied to the original time series, as described in Equation 4.

$$(equation\ 4) \quad Q_i' = (Q_i + r_i + k)\frac{\sum_{i=1}^{n} \quad Q}{\sum_{i=1}^{n} \quad Q_i'}$$

Where Q_i' is the resulting Q for the day, Q_i is the observed discharge at that time, r_i is the resampled residual for that day, *k* is a constant to prevent zero flow days, and the ratio term is a hold-factor which adjusts the total water yield each generated year the water year of the input time series.

The same modeling method was then applied to generate stormflow - and baseflow-dominated discharge series, as described in Equations 5 and 6 respectively.

$$(Equation\ 5) \quad Q_i' = (Q_i^{1.5} + r_i + k)\frac{\sum_{i=1}^{n} Q}{\sum_{i=1}^{n} Q_i'}$$

$$(Equation\ 6) \quad Q_i' = (Q_i^{0.9} + r_i + k)\frac{\sum_{i=1}^{n} Q}{\sum_{i=1}^{n} Q_i'}$$

The exponent applied to Q_i attenuates the flow, creating a larger or smaller stormflow signal. The baseflow time series was then attenuated with a moving average to reduce noise using the “zoo” R package’s “rollmean” function with a *k* value of 10 (Zeileis and Grothendieck, 2005).

We repeated this process 100 times to create a range of potential site-years under each hydrologic regime. An example of a single run is illustrated in Figure 7.

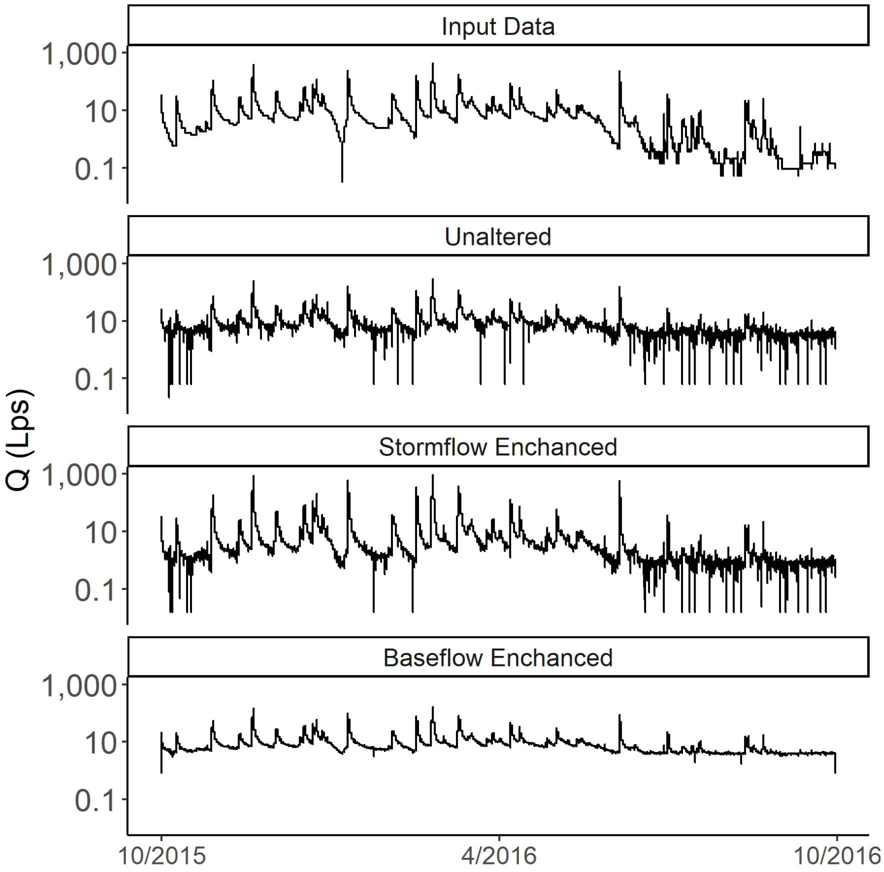

Figure 7: Examples of the hydrology time series simulated for this study. The raw, input data (top) was used to fit an ARIMA model. The residuals of this model fit were then resampled and added to the original series to create the three modeled time series. This process was repeated 100 times. The unaltered time series only has reshuffled residuals. The stormflow enhanced series has been transformed to have a ~15% increase in quickflow. The baseflow enhanced

time series had a ~15% reduction in quickflow. Note that each time series has the same total flux of water, only the timing of the delivery changes.

We used the “EcoHydRology” R package”s “BaseflowSeparation” function to determine proportion of quickflow to total flow for each time series (Fuka et al., 2014). The unaltered time series had ~20% quickflow, the stormflow time series had ~35% quickflow, and the baseflow time series had ~5% quickflow.

2.3.2 C:Q regimes

To generate chemostatic time series of stream chemistry, we randomly sampled a normal distribution of points with a mean of 2 mg/L and a standard deviation of 0.1 mg/L for each day in the generated streamflow time series. Likewise, to generate our no-pattern time series we randomly sampled a normal distribution of points with a mean of 2 mg/L and a standard deviation of 0.5 mg/L for each day in the generated streamflow time series.

To generate time series of enriching stream chemistry, we applied Equation 7 to our generated streamflow time series.

$$(equation\ 7) \quad C_i' = 10^{log10(Q_i')-1}$$

Similarly, we applied Equation 8 to generate time series with diluting chemistry.

$$(equation\ 8) \quad C_i' = 10^{log10(-Q_i')+1.5}$$

We then added error to the enriching and diluting time series by taking the product of each C_i' and an error factor, randomly drawn for each day from a normal distribution with a mean of 1 and a standard deviation of 0.1.

We applied each of these methods to each hydrology time series generated in the previous section. An example of the resulting C:Q relationships formed are illustrated in Figure 8.

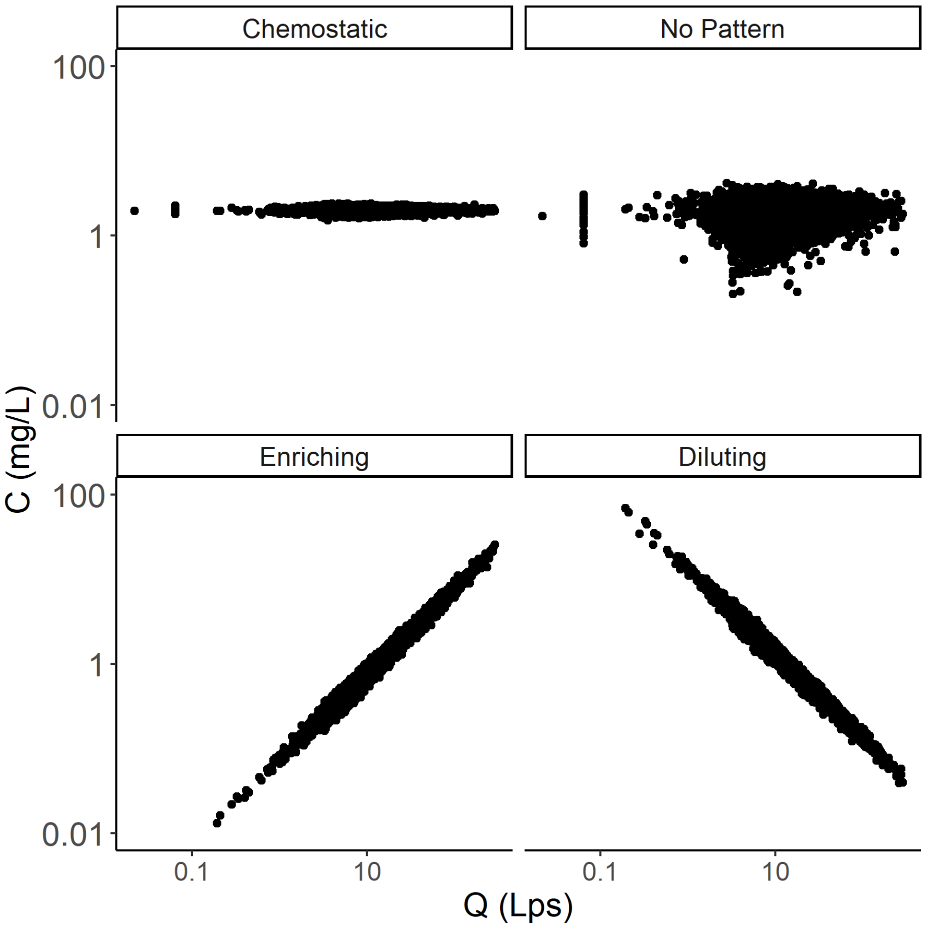

Figure 8: An example of the C:Q regimes created using synthetic hydrology data. Each hydrology time series from the previous section was used to create four concentration time series. The resulting C:Q relationships from applying each of our chemistry generation equations (chemostatic, no pattern, enriching, and diluting) to one run of the simulated, non-baseflow modified, modeled time series is represented here.

To assess method accuracy, we calculated “true” flux for each synthetic solute year using the full synthetic time series and the composite method (as recommended in Aulenbach et al. 2016) to estimate over incomplete days. The composite method is ideal for this, as the high density of the data yields highly autocorrelated residuals (Aulenbach et al., 2016). We coarsened each synthetic time series of concentration and discharge to the weekly, biweekly, and monthly time steps. Then, we applied each of our four flux methods to each coarsened time series and compared the generated estimates to our true annual flux.

2.4 Generating estimates for 93 watersheds

MacroSheds is a synthesis dataset of long-term biogeochemical, hydroclimatic, and geospatial data from small watershed ecosystem studies. The full dataset is available to the public at https://portal.edirepository.org/nis/mapbrowse?scope=edi&identifier=1262&revision=1 and the latest version is linked at macrosheds.org. The dataset includes harmonized data from 93 federally funded watershed studies from across the United States (Vlah et al., 2023). To provide flux estimates to the broader community, we applied all four of our flux estimation methods, along with a simple average, as described in Aulenbach et al. 2016, to the MacroSheds dataset. Methods were applied individually to solutes at the site-year level. A simplified application of the Aulenbach et al. 2016 decision framework was applied to each site-year to give a recommended method. Our method fit a log-log linear model between concentration and discharge for the solute year. We then calculated the autocorrelation of concentration values, the R-squared value of the model and the autocorrelation of the model residuals. Next, we sorted each solute year using a cutoff of 0.30 for the model R-squared. Years with a model R-squared greater than or equal to 0.30 were designated as having “strong to moderate” fits, while years with R-squared lower than 0.30 were designated as having “weak to nonexistent” fits. Solute years with a “strong to moderate” fit and autocorrelated residuals (>=0.20) are recommended to use the composite method, and those without or with weak autocorrelation (<0.20) were recommended to use the rating method. Solute years with a “weak to nonexistent” C:Q fits and autocorrelated concentrations (>=0.20) were recommended to use linear interpolation, while those without or with weak autocorrelation of concentrations (<0.20) were recommended to use a simple average. Only sites with 85% or more days of discharge coverage and at least one chemistry sample per water year quarter were used to calculate loads.

3 Results

3.1 Method sensitivity to actual high-frequency data

Results from the data coarsening experiments at Hubbard Brook are represented in Figures 9 and 10. Generally, all methods struggled to accurately estimate nitrate load as data became increasingly coarse.

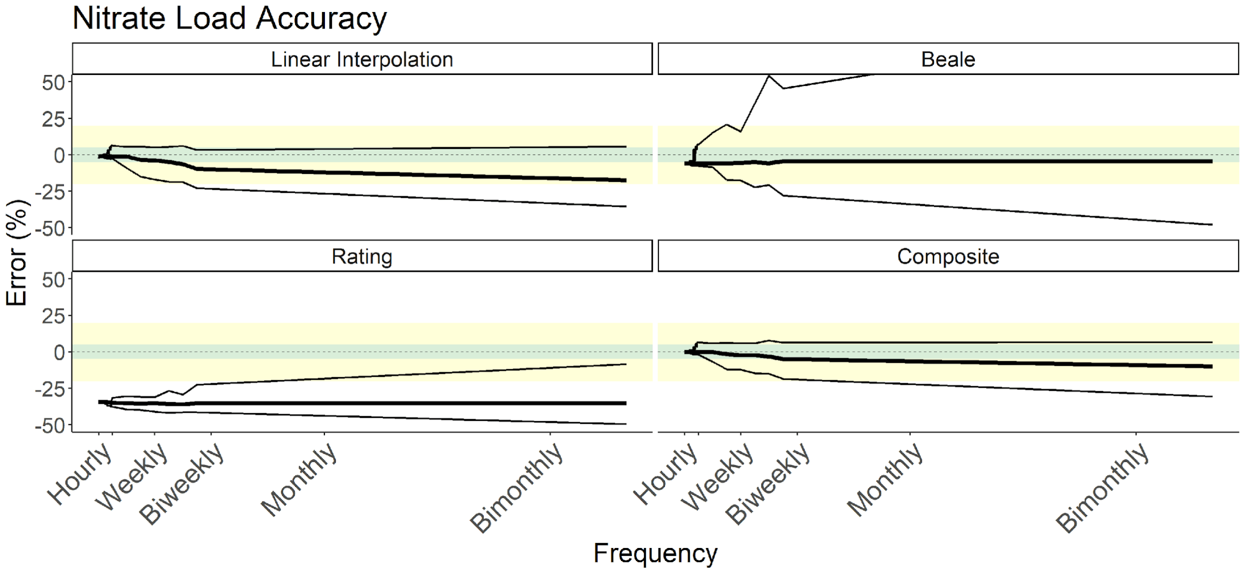

Figure 9: Results from applying four different load estimation methods to artificially coarsened, high-frequency sensor time series of nitrate-N taken from Hubbard Brook Watershed 3. Percent error of estimated annual load from the

"true" load, calculated using the full time series, is on the y-axis. Coarsened data frequency is on the x-axis. The solid line indicates the median error for that frequency, with thin lines indicating minimum and maximum. Error within 5% of truth is shaded in green and within 20% in yellow. Note that the linear interpolation and composite methods perform best for this year of data. However, all methods performed poorly with sample frequencies coarser than daily.

Overall, the composite method performed best for this solute year, with accurate median estimates and a relatively low range of potential estimates. Linear interpolation performed comparably with sub-weekly data. When data frequency was high, both concentrations and model residuals were highly autocorrelated. As expected, under these conditions, methods that leverage autocorrelation, such as linear interpolation and composite, outperform all others. As sampling becomes more infrequent and the autocorrelation of concentrations breaks down, linear interpolation becomes less accurate than the composite method. This is contrary to the guidelines suggested by Aulenbach et al. (2016), which–based on the low (~0.07) R-squared of the model fit–would recommend linear interpolation or averaging. Looking at the C:Q relationship presented in Figure 3, we see that while the fit is poor, there is still an informative, enriching pattern that the composite method leverages for more accurate interpolation than simple linear interpolation. Temporally variable changes in C:Q relationships and storm responses have been found to reduce load estimate quality (Fazekas et al., 2021), which the composite method's residual corrections help to mitagate. While relying on the fit alone (the rating method) produces uniformly poor results, including residual corrections greatly improves model performance. It should be noted that the Aulenbach et al. 2016 selection method does still select a defensible method (linear interpolation or averaging, depending on sampling frequency), without the benefit of referring to the high-frequency time series.

There was a tight relationship between Ca and specific conductance at Watershed 3 during the 2016 water year. The fitted linear model had an R-squared of 0.92 and a slope of 0.0063.

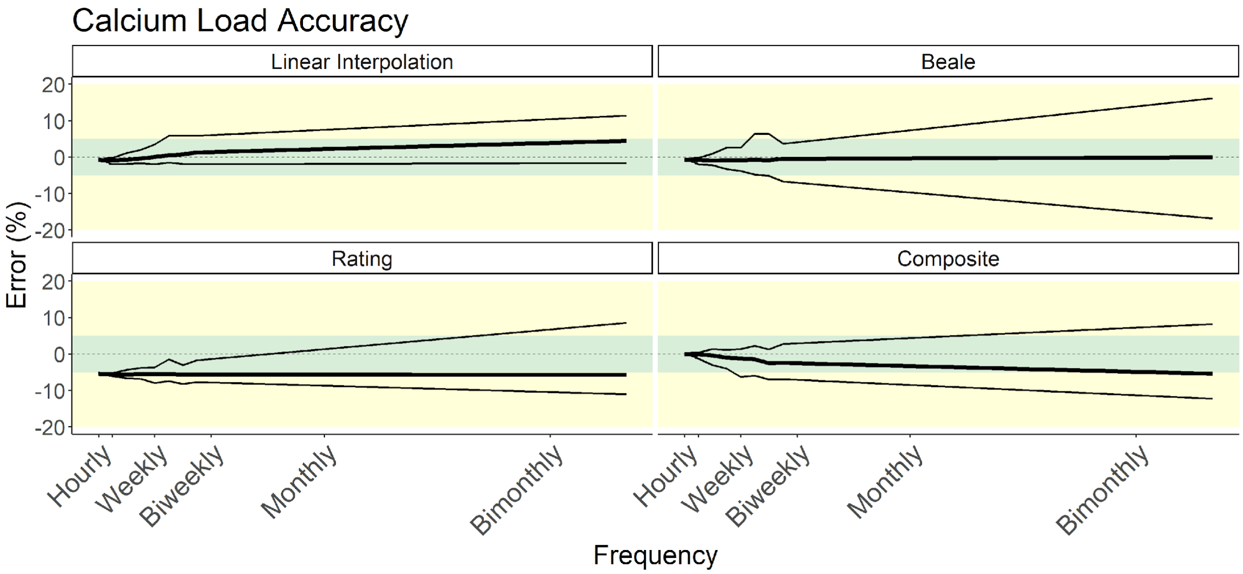

Figure 10: Results from applying four different load estimation methods to artificially coarsened, high-frequency sensor time series of calcium (estimated from specific conductivity) taken from Hubbard Brook watershed 3. Percent error of estimated annual load from the "true" load, calculated using the full time series, is on the y-axis. Coarsened data frequency is on the x-axis. The solid line indicates the median error for that frequency, with thin lines indicating

minimum and maximum. Error within 5% of truth is shaded in green and within 20% in yellow. Note that linear interpolation generally performs best for this year of data. All methods, excluding the rating method, provided good quality estimates with weekly or finer data frequency.

Generally, linear interpolation performed best for this solute year, regardless of sampling frequency. This is again contrary to what the Aulenbach et al. 2016 method would suggest, which - based on the strong (~0.79) R-squared of the model fit - would suggest the rating or composite method. Again, referring to the C:Q relationship from Figure 3, we see that there is an enriching trend at high flows that is not captured in our simple model. This is likely what drives the underprediction shown in the composite method results. The range of estimates produced by the rating and composite methods was higher across coarse sampling frequencies than from linear interpolation. As with the NO_3-N results, the Aulenbach et al. 2016 suggested method does not concur with our analysis but does select a defensible method.

For both NO_3-N and Ca, the Beale ratio estimator produced unbiased median estimates from coarse sampling frequencies at the cost of high estimate variance. Due to the underlying assumption of covariance of discharge variance and concentration variance the method performed more accurately on the Ca time series than the NO_3-N time series. Users with coarse data interested primarily in reducing bias, rather than error, should consider using Beale derived estimates.

Comparing the results from NO_3-N and Ca across methods, we see that using C:Q model fits is a flawed, but useful tool in method selection. While the NO_3-N C:Q fit is weak, it provides a useful trend that improves on simply linearly interpolating points. In contrast, even though the Ca time series has a strong C:Q model fit, the model does not accurately describe the true nature of the relationship, chronically biasing our results low. Both cases illustrate the complexity and difficulty of making method selection choices without the benefit of “true” estimates with which to compare.

Compared to Watershed 3 at Hubbard Brook, the Upper Hafren dataset from Plynlimon has much less variable stream chemistry. As shown in Figures 11 and 12, this lower variance yielded more precise estimates across all methods. Similar to at Hubbard Brook, NO_{3-}N load at Plynlimon was best estimated by the linear interpolation and composite methods.

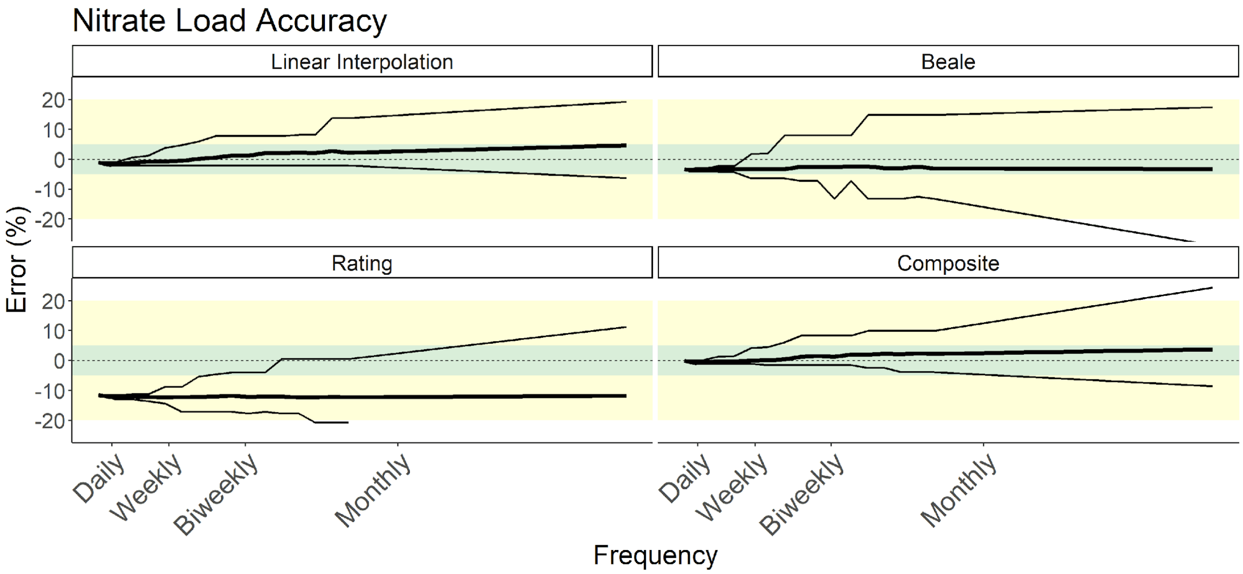

Figure 11 Results from applying four different load estimation methods to artificially coarsened, high-frequency sensor time series of nitrate-N taken from Plynlimon Research Catchments' Upper Hafren site. Percent error of estimated annual load from the "true" load, calculated using the full time series, is on the y-axis. Coarsened data frequency is on the x-axis. The solid line indicates the median error for that frequency, with thin lines indicating minimum and maximum. Error within 5% of truth is shaded in green and within 20% in yellow. Note that the linear interpolation and composite methods perform best for this year of data.

Comparing Figures 10 and 12 highlights the diminishing importance having an informative C:Q relationship to generate load estimates as chemistry becomes less dynamic. When stream chemistry variation is low, even a less informative C:Q relationship can be leveraged to produce more accurate load estimates. In this case, having a complete understanding of water flux is enough to produce accurate Ca load estimates even with monthly chemistry data.

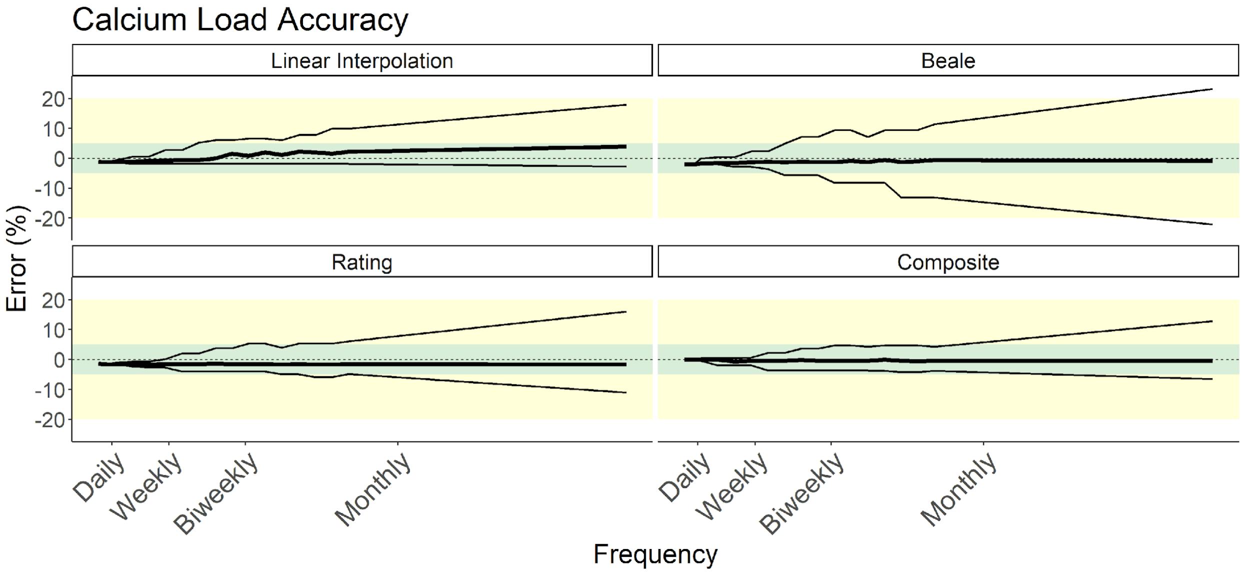

Figure 12: Results from applying four different load estimation methods to artificially coarsened, high-frequency sensor time series of Calcium, taken from Plynlimon Research Catchments' Upper Hafren site. Percent error of estimated annual load from the "true" load, calculated using the full time series, is on the y-axis. Coarsened data frequency is on the x-axis. The solid line indicates the median error for that frequency, with thin lines indicating

minimum and maximum. Error within 5% of truth is shaded in green and within 20% in yellow. Note that the composite method performed exceptionally well here, producing reliable load estimates even with monthly data.

3.2 Synthetic time series

The results for our synthetic time series experiments are shown in Figure 13. Loads of chemostatic solutes were universally the easiest to predict. Generally, solutes with no obvious relationship to discharge were the most difficult to predict, regardless of method. All methods degraded with increasingly coarsened data. A full table of the results from Figure 13 are available in supplementary_table_1.csv.

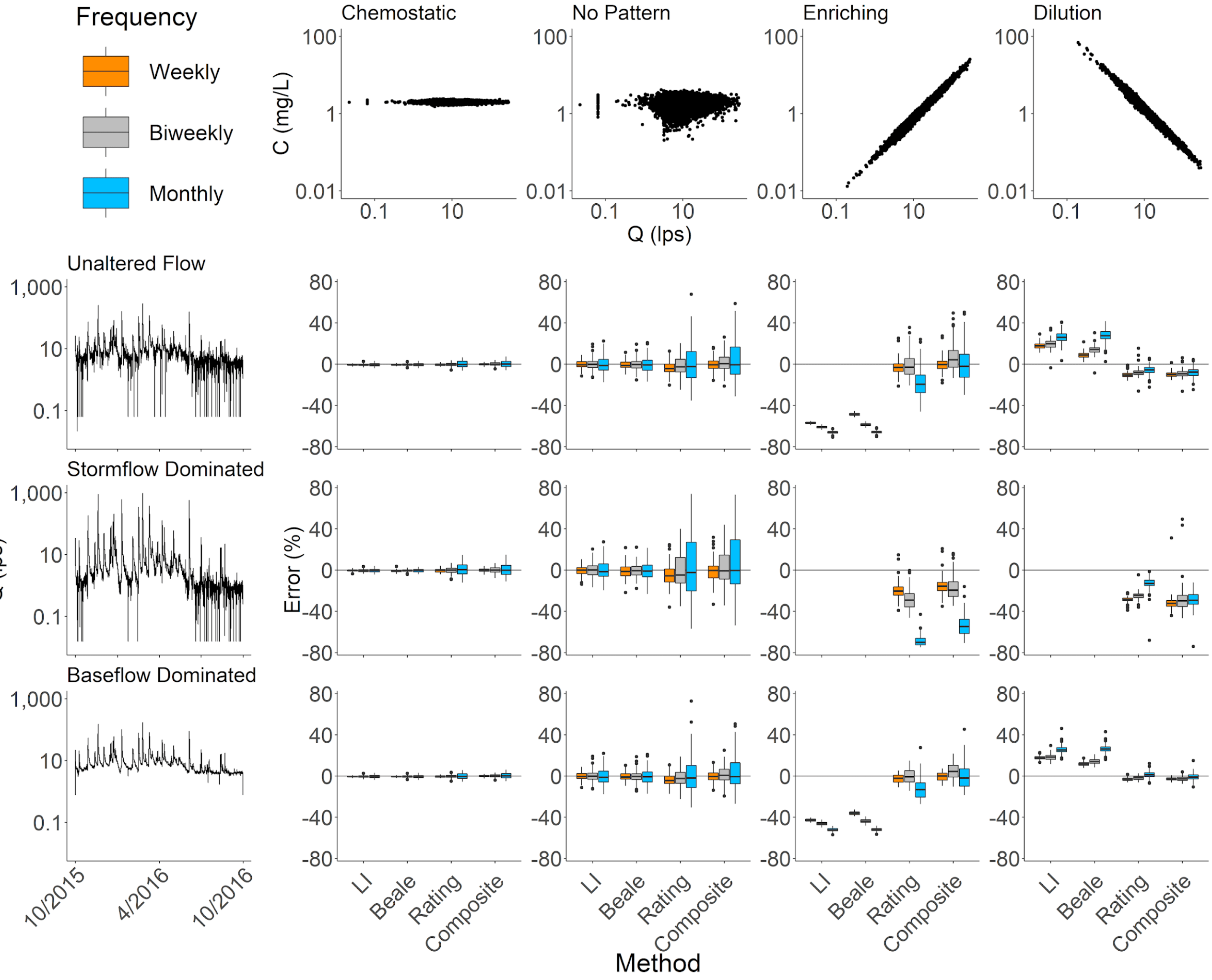

Figure 13: Load estimation error across chemodynamic and hydraulic regimes, and four estimation methods. Load estimates were generated from simulated data. The top row of plots describes the concentration discharge relationship. The side column of plots describes the hydrologic regime. The grid of boxplots is the result of applying the four estimation methods across each combination of concentration discharge relationship and hydrologic regime. Linear interpolation (LI) and Beale estimates from enriching or diluting time series under stormflow conditions, and some outlier values (beyond 1.5 interquartile ranges) excluded for readability. A table of summary values for all methods, hydrologic regimes, and C:Q relationships is available in supplementary_table_1.csv. Note that no-pattern

consistently shows high error, while chemostatic (generally) and enriching (with composite method and high data frequency) give low error.

3.3 Application to MacroSheds dataset

Applying our methods to the MacroSheds dataset generated 16,489 site-years of data across 93 sites and 112 solutes. Distributions of annual solute loads of nitrate (as nitrogen) and calcium are shown in Figure 14. The complete load calculations from each solute and site-year of MacroSheds data are available at 10.6084/m9.figshare.24975504.

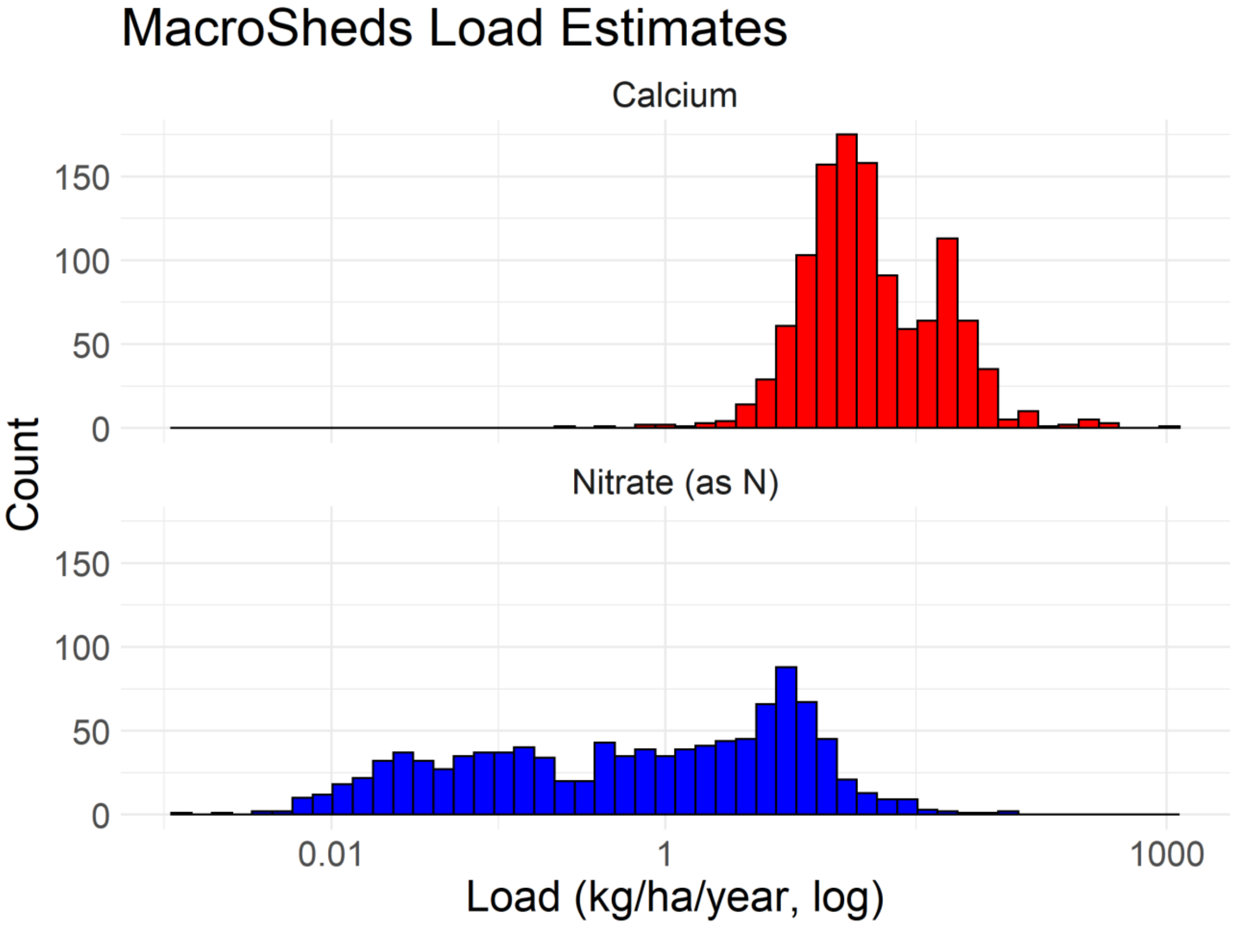

Figure 14: A histogram of annual load estimates present in the MacroSheds dataset. The complete load calculations from each solute and site-year of MacroSheds data are available at 10.6084/m9.figshare.24975504*.*

4 Discussion

4.1 Insights on load estimation uncertainty

Our results from both sets of experiments corroborate what Aulenbach and others (2016) observed in their assessment of load estimation methods in small watersheds. Namely that densely sampled data is usually best estimated by leveraging highly autocorrelated rating residuals (composite method) or highly autocorrelated concentrations (linear interpolation), and that informative C:Q relationships should be used where possible. Using our simplified application of Aulenbach and others' (2016) method selection procedure yields a defensible selection of load estimation methods. As shown in Figure 10, other single methods also provide reasonable estimates. However, it would be difficult to confidently assess their relative effectiveness without a sensor derived "truth" for comparison.

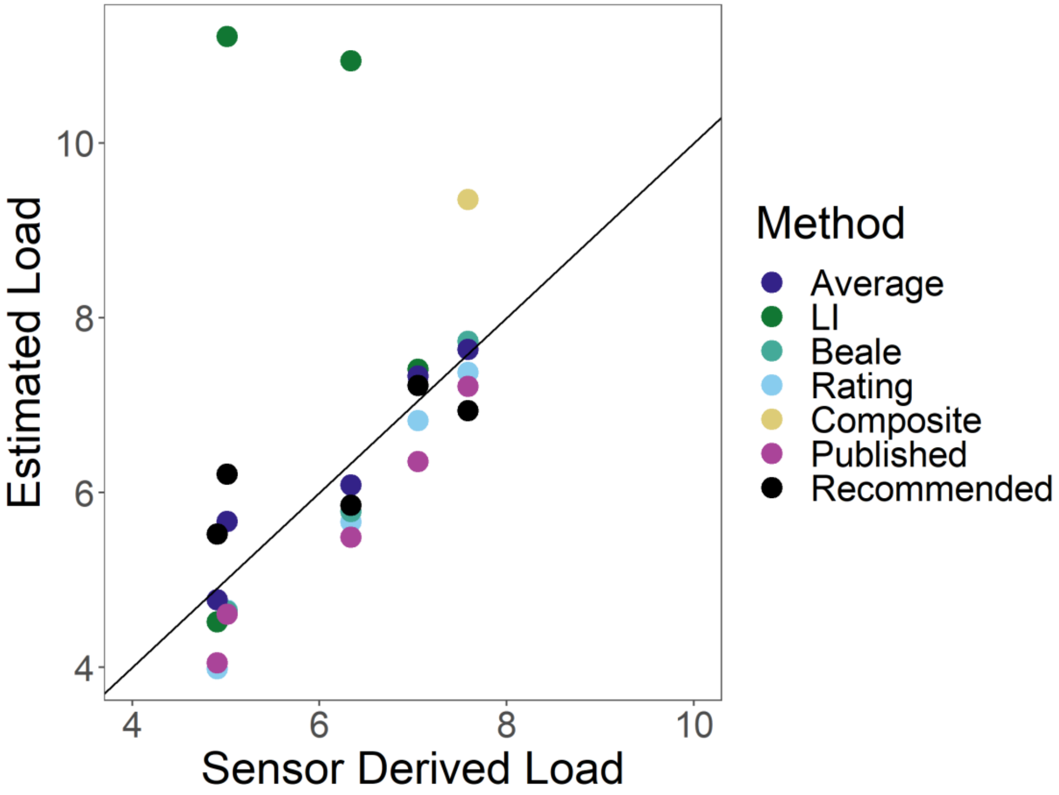

Figure 15: A comparison between load estimation methods made with data from watershed 3 at Hubbard Brook Experimental Forest. Sensor derived load is calculated using 15-minute specific conductance data, converted to calcium concentrations via regression, with the composite method applied. Estimated loads were calculated using biweekly, discrete grab samples. Values published from Hubbard Brook are in magenta. The estimate recommended by our simplified application of Aulenbach et al. 2016 is in black. A 1:1 line is plotted in black. The recommended method yielded an R-squared of 0.63.

The general agreement between the recommended method estimates and sensor derived load estimates in Figure 15 is a reaffirming case study that the Aulenbach 2016 decision flowchart sensibly chooses from among the best load estimation methods, without requiring the use of outside data (R-squared of 0.63 between estimated and sensor derived loads).

Our results also reinforce previous findings that – generally – when there is not a strong C:Q relationship, users should rely on linear interpolation or averaging methods. Under chemostatic or no-pattern C:Q relationships, linear interpolation and the Beale estimator outperform the rating and composite methods, regardless of hydrologic regime (Fig. 15). Consider a load estimate from a solute with no underlying C:Q pattern, stormflow dominated hydrology, and monthly sampling. Our experiment shows linear interpolation yields a mean error of -0.43%

(95% confidence interval of -19.11% to 18.25%). Meanwhile, the composite method yields a mean error of 8.44% (95% confidence interval of -55.67% to 72.55%).

When there is a strong C:Q relationship, researchers should use a method that leverages that relationship. Our results in Figure 13 confirm that when the C:Q relationship is stable and effectively modeled, C:Q informed methods tend to outperform others. Under such conditions, the rating and composite methods dramatically outperform linear interpolation or the Beale ratio estimator. For example, a load estimate from a solute with a diluting C:Q relationship, stormflow dominated hydrology, and monthly sampling has a mean error of -13.43% (95% confidence interval of –27.44% to 0.58%), while linear interpolation yields a mean error of 119.57% (95% confidence interval of 85.47% to 153.67%).

Results from the data coarsening experiment give nuance to the synthetic time series experiments. Methods applied less cleanly to the C:Q relationship in the coarsened Ca time series (Figure 10) than in our synthetic time series testing. The C:Q relationship present in the calcium time series (Figure 3) gives a high R-squared of 0.79 and a low slope of -0.12. While it produces low mean error, linear interpolation (the best available method) still produced estimates that overpredicted load by as much as 11% at monthly sampling frequencies (Figure 10). In the synthetic time series experiment, we expected much smaller error given C:Q relationship and sampling frequency, with maximum error of 3.34%. This ~7% absolute difference shows a limitation of our synthetic time series analysis: the variance of the underlying time series (both in streamflow and chemistry) has a scaling effect on uncertainty. The calcium time series has a standard deviation of 0.23 mg/L, while the synthetic, chemostatic time series has a standard deviation of 0.1 mg/L.

Considering the results of the synthetic time series and comparing our Hubbard Brook (Figures 9 and 10) results to our Plynlimon results (Figures 11 and 12), highlights an important conclusion for macroscale scientists interested in load estimation: if variance in solute chemistry is sufficiently low, knowing water flux is enough to accurately estimate solute loads. However, doing so with confidence requires high-frequency data or an explanatory, well-defined C:Q relationship.

4.2 Challenges for macroscale science with load estimates

The results of our efforts highlight several important challenges for scientists working with load estimates. First, method selection is critically important for accurate load estimation. The results of our data coarsening experiment (Figures 9-12) clearly show that–especially at coarse sampling intervals–method selection may influence estimate accuracy by up to ~50 percent. Therefore, load estimates provided without documentation of the underlying methodology should be treated with great caution, or recomputed from chemistry and discharge using tools like the “RiverLoad” and “MacroSheds” R packages. By extension, these findings suggest that load estimate data providers should always release underlying concentration and discharge data.

Our experiments suggest that non-chemostatic load estimates should be used with careful consideration, especially in stormflow dominated areas or at weekly to monthly sampling frequencies. While optimal methods may produce accurate estimates, even at these frequencies the extent of possible error observed over 100 runs was still problematic. For example, under a stormflow dominated hydrologic regime, with no C:Q relationship, and weekly sampling, linear interpolation (the best performing method) gave a mean error of -0.2%, but the 95% confidence interval spanned from -10.59 to 10.92%.

Finally, determining the direction of the C:Q relationship alone is not enough to assign confidence to underlying load estimates generated from that solute time series. Confidence in that determination, with respect to the full range of possible flow conditions, must also be assessed. A site that biases its collection towards baseflow days (which is common with non-event supplemented sampling) may erroneously conclude that their enriching solute"s C:Q relationship is chemostatic or has no pattern (Aulenbach et al., 2016). A user making this conclusion would be tempted to use linear interpolation or the Beale ratio estimator to reduce both error and bias. This yields load estimates that heavily underestimate true load for the year. The importance of a well-defined C:Q relationship is also evident when considering the effect of diluting conditions in Figure 13. Moving from weekly to monthly sampling frequencies counterintuitively decreases bias, while increasing variance. This is likely due to an overfitting of the rating model to baseflow points decreasing rating accuracy at low flows. Results from the data coarsening experiments suggest this trend would reverse at high sampling frequencies. We see a similar effect in the enriching time series, where biweekly sampling shows less bias than weekly sampling. These data suggest that simply “fitting and forgetting” a rating model is not enough to generate the best possible load estimates.

Put simply, truly assessing confidence in a load estimate requires an assessment of confidence in both the C:Q relationship itself and an assessment of confidence in having the entire C:Q relationship. All the challenges delineated here point to a clear need for more long-term, high-frequency records of stream chemistry, as well as the potential utility of using sensor data to help assess the accuracy of longer term low-frequency sampling efforts. While it may be eventually possible to estimate loads accurately from sparse records using machine learning or other emerging computing methods, currently there is no substitute for high quality, frequent observations.

4.3 Framework for considering load estimates

The challenges from the previous section – ensuring proper method selection, overcoming low all-method accuracy in highly variable systems, and assessing confidence in knowledge of the entire C:Q relationship at a site – are easy to identify, but difficult to solve, especially for scientists interested in cross-site comparisons. While other researchers can rely on an intimate understanding of the history and disturbances at each site to inform their confidence in estimates, macroscale scientists must rely only on the data provided from previous studies. Important site history narratives and key assumptions may be difficult to locate in publications and metadata, if they are recorded at all. Cultivating a rich and well-informed understanding of every individual site is often not feasible, and depending on publicly accessible documentation, may be impossible. However, excluding good data from a synthesis effort is wasteful and limits the power of analyses and the scope at which synthesis can be performed. Therefore, it is necessary to develop a framework of assessing load estimate confidence using only the site records themselves.

We propose the framework presented in Figure 16 to classify site years by ease of creating high-confidence load estimates for cross-site comparisons. As C:Q relationships have been shown to change over time (Kirchner et al., 2004; Godsey et al., 2019, Fazekas et al., 2021) and this framework is built on the assumption that the user does not have intimate site knowledge, each site-year is assessed independently. Our framework also assumes users have a complete flow record for the site and that they have chosen the most appropriate estimation method for each solute as described in Aulenbach et al. 2016.

With effort, this framework could be greatly improved by grouping data and method selection across multiple solute years with known, stable chemodynamics. Doing so would leverage the power of long-term data and allow for much more well-informed C:Q models - generating higher confidence estimates. However, doing so requires the user to have a reliable method to assess the stability of C:Q relationships within individual solutes and to develop methods to handle transitions between regimes. Currently, the authors are not aware of any such methods or tools.

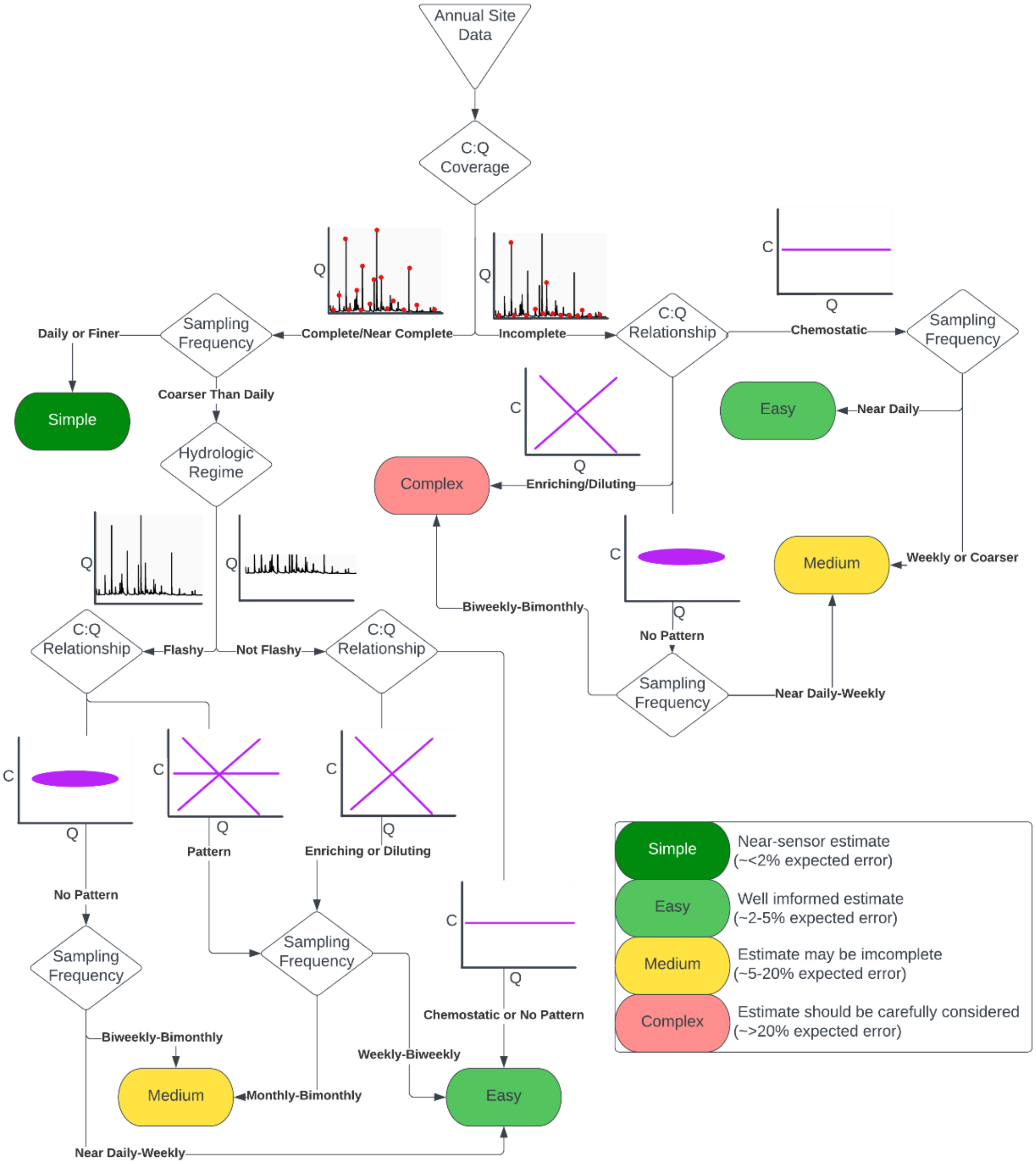

Figure 16: A conceptual flowchart for classifying confidence in load estimates, assuming a complete discharge time series and method selection as described in Aulenbach et al. 2016. C:Q coverage refers to the completeness of concentration observations relative to possible flow conditions. Sampling frequency is the interval at which chemistry is sampled at the site. Hydrologic regime is either defined as flashy (stormflow dominated) or not flashy (baseflow dominated). Site-years with total C:Q coverage and high sampling frequency make it 'simple' to produce quality estimates. Non-chemostatic site-years with incomplete C:Q coverage have 'medium' difficulty producing quality estimates or require 'complex' methods and considerations.

Data binned as “simple” or “easy” is likely suitable for inter-site comparisons at large scales. Data binned as “fair” should only be used for limited applications. For example, “medium” rated estimates could be used in aggregated regional estimates of weathering rates. Data binned as “complex” should not be used without the user learning more about the site. It is possible that “complex” rated estimates have lower error than expected, especially if the site they are derived from has been long-running and uses targeted sampling. It should be noted that while error ranges are presented for each category, they are a qualitative assessment guided by the results of our experiments. Further work will be required to truly constrain expected error for such a complex problem.

While it is beyond the scope of this effort to delineate exact boundaries for each branch in Figure 18, we fully expect to be able to do so robustly as more sensor data becomes available in a variety of small watershed systems. With adequate, widely distributed sensor data, we expect the factors identified in our framework could be used to sort load estimates by relative quality. C:Q coverage can be systematically assessed by creating a ratio of sampled flows to observed flows over the year. Sampling frequency can be easily evaluated from the stream chemistry record. The shape of the C:Q relationship can be determined by fitting a log-log simple linear model between solute concentrations and discharge at the site and assessing the slope, R-squared, and residuals of the resulting fit. Chemostatic and no pattern time series could be differentiated using a ratio of concentration time series” standard deviation to its mean. Hydrologic regime can be assessed using the Richards-Baker flashiness index (Baker et al., 2004) or using the baseflow-quickflow separation method used in this study. While the field gathers more sensor data, we encourage the larger macroscale watershed science community to continue to test and model the effects of these variables on load estimation. Developing robust, transparent, and automatable approaches to inform load estimation methods around these decision points will facilitate scalable and reliable cross-site load comparisons in macroscale science.

While many of our proposed, fundamental questions remain unresolved, we hope this analysis can be used as a roadmap to focus the community on critical knowledge gaps in cross-site load comparisons. Namely, we implore future research to continue to work on developing robust breakpoints for the decision points presented in Figure 18 (and discussed above) and on assessing the stability and applicability of solute chemodynamics across time. Doing so in robust, automatable ways will unlock macroscale science efforts that rely on load estimates and empower synthesis scientists to delve into the rich library of datasets available to them from past studies.

5 Works Cited

Aulenbach, B. T., & Hooper, R. P. (2006). The composite method: An improved method for stream-water solute load estimation. *Hydrological Processes*, *20*(14), 3029–3047. https://doi.org/10.1002/hyp.6147

Baker, D. B., Richards, R. P., Loftus, T. T., & Kramer, J. W. (2004). *A NEW FLASHINESS INDEX: CHARACTERISTICS AND APPLICATIONS TO MIDWESTERN RIVERS AND STREAMS*. *44883*(03095), 503–522.

Beale, E. (1962). Some uses of computers in operational research. *Industrielle Organisation*, *31*, 51–52.

Bormann, A. F. H., Likens, G. E., Fisher, D. W., & Pierce, R. S. (1968). Nutrient Loss Accelerated by Clear-Cutting of a Forest Ecosystem. *Science*, *159*(3817), 882–884.

Buso, D. C., Likens, G. E., & Eaton, J. S. (2000). *Chemistry of Precipitation , Streamwater , and Lakewater from the Hubbard Brook Ecosystem Study : A Record of Sampling Protocols and Analytical Procedures*. 52. http://www.fs.fed.us/ne/newtown_square/publications/technical_reports/pdfs/2000/gtrne275.pdf

Charles Gene Crawford. (1996). ESTIMATING MEAN CONSTITUENT LOADS IN RIVERS BY THE RATING-CURVE AND FLOW-DURATION, RATING-CURVE METHODS. In *School of Public and Environmental Affairs Indiana University*.

Coombs, J. S., & Melack, J. M. (2013). Initial impacts of a wildfire on hydrology and suspended sediment and nutrient export in California chaparral watersheds. *Hydrological Processes*, *27*(26), 3842–3851. https://doi.org/10.1002/HYP.9508

Crawford, J. T., Hinckley, E. L. S., Litaor, M. I., Brahney, J., & Neff, J. C. (2019). Evidence for accelerated weathering and sulfate export in high alpine environments. *Environmental Research Letters*, *14*(12). https://doi.org/10.1088/1748-9326/ab5d9c

Dodds, W. K., Bouska, W. W., Eitzmann, J. L., Pilger, T. J., Pitts, K. L., Riley, A. J., Schloesser, J. T., & Thornbrugh, D. J. (2009). Eutrophication of U. S. freshwaters: Analysis of potential economic damages. *Environmental Science and Technology*, *43*(1), 12–19. https://doi.org/10.1021/es801217q

Fazekas, H. M., Mcdowell, W. H., Shanley, J. B., & Wymore, A. S. (2021). Climate Variability Drives Watersheds Along a Transporter-Transformer Continuum. *Geophysical Research Letters*. https://doi.org/10.1029/2021GL094050

Fuka, D., Walter, M., Archibald, J., Steenhuis, T., & Easton, Z. (2014). *A community modeling foundation for Eco-Hydrology* (0.4.12; p. 5). R. do

Godsey, S. E., Hartmann, J., & Kirchner, J. W. (2019). Catchment chemostasis revisited: Water quality responds differently to variations in weather and climate. *Hydrological Processes*, *33*(24), 3056–3069. https://doi.org/10.1002/hyp.13554

Godsey, S. E., Kirchner, J. W., & Clow, D. W. (2009). Concentration-discharge relationships reflect chemostatic characteristics of US catchments. *Hydrological Processes*, *23*(13), 1844–1864. https://doi.org/10.1002/hyp.7315

Halliday, S. J., Wade, A. J., Skeffington, R. A., Neal, C., Reynolds, B., Rowland, P., Neal, M., & Norris, D. (2012). An analysis of long-term trends, seasonality and short-term dynamics in water quality data from Plynlimon, Wales. *Science of the Total Environment*, *434*, 186–200. https://doi.org/10.1016/j.scitotenv.2011.10.052

Hirsch, R. M., Moyer, D. L., & Archfield, S. A. (2010). Weighted regressions on time, discharge, and season (WRTDS), with an application to chesapeake bay river inputs. *Journal of the American Water Resources Association*, *46*(5), 857–880. https://doi.org/10.1111/j.1752-1688.2010.00482.x

Kirchner, J. W., Feng, X., Neal, C., & Robson, A. J. (2004). The fine structure of water-quality dynamics: The (high-frequency) wave of the future. *Hydrological Processes*, *18*(7), 1353–1359. https://doi.org/10.1002/hyp.5537

Koger, J. M., Newman, B. D., & Goering, T. J. (2018). Chemostatic behaviour of major ions and contaminants in a semiarid spring and stream system near Los Alamos, NM, USA. *Hydrological Processes*, *32*(11), 1709–1716. https://doi.org/10.1002/hyp.11624

Lee, C. J., Hirsch, R. M., & Crawford, C. G. (2019). *An Evaluation of Methods for Computing Annual Water-Quality Loads*.

Likens, G. E., Bormann, F. H., Johnson, N. M., Fisher, D. W., & Pierce, R. S. (1970). Effects of Forest Cutting and Herbicide Treatment on Nutrient Budgets in the Hubbard Brook Watershed-Ecosystem. *Ecological Monographs*, *40*(1), 23–47. https://doi.org/10.2307/1942440

Likens, G. E., & Buso, D. C. (2006). Variation in streamwater chemistry throughout the Hubbard Brook Valley. *Biogeochemistry*, *78*(1), 1–30. https://doi.org/10.1007/s10533-005-2024-2

Maher, K., & Chamberlain, C. P. (2014). Hydrologic regulation of chemical weathering and the geologic. *Science*, *343*(6178), 1502–1504. https://doi.org/10.1126/science.1250770

McDowell, W. H., McDowell, W. G., Potter, J. D., & Ramírez, A. (2019). Nutrient export and elemental stoichiometry in an urban tropical river. *Ecological Applications*, *29*(2), 1–15. https://doi.org/10.1002/eap.1839

Meals, D. W., Richards, R. P., & Dressing, S. A. (2013). Pollutant Load Estimation for Water Quality Monitoring Projects. *Tech Notes 8*, *1*(1), 21. https://www.epa.gov/ polluted-runoff-nonpoint-source-pollution/nonpoint-source-monitoring- technical-notes

Moatar, F., Abbott, B. W., Minaudo, C., Curie, F., & Pinay, G. (n.d.). *Elemental properties, hydrology, and biology interact to shape concentration-discharge curves for carbon, nutrients, sediment, and major ions*. https://doi.org/10.1002/2016WR019635

Neal C.;Kirchner, J. ;Reynold. B. (2013). *Plynlimon research catchment high-frequency hydrochemistry data*. NERC Environmental Information Data Centre. https://doi.org/10.5285/551a10ae-b8ed-4ebd-ab38-033dd597a374

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T., & Duan, Q. (2015). Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, *19*(1), 209–223. https://doi.org/10.5194/hess-19-209-2015

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

Petrone, K. C., Jones, J. B., Hinzman, L. D., & Boone, R. D. (2006). Seasonal export of carbon, nitrogen, and major solutes from Alaskan catchments with discontinuous permafrost. *Journal of Geophysical Research: Biogeosciences*, *111*(2), 1–13. https://doi.org/10.1029/2005JG000055

Quilbé, R., Rousseau, A. N., Lafrance, P., Leclerc, J., & Amrani, M. (2006). Selecting a pesticide fate model at the watershed scale using a multi-criteria analysis. *Water Quality Research Journal of Canada*, *41*(3), 283–295. https://doi.org/10.2166/wqrj.2006.032

Richards, R. P., & Holloway, J. (1987). Monte Carlo studies of sampling strategies for estimating tributary loads. *Water Resources Research*, *23*(10), 1939–1948. https://doi.org/10.1029/WR023i010p01939

Schilling, K. E., Jones, C. S., Wolter, C. F., Liang, X., Zhang, Y. K., Seeman, A., Isenhart, T., Schnoebelen, D., & Skopec, M. (2017). Variability of nitrate-nitrogen load estimation results will make quantifying load reduction strategies difficult in Iowa. *Journal of Soil and Water Conservation*, *72*(4), 317–325. https://doi.org/10.2489/jswc.72.4.317

See, C. R., Green, M. B., Yanai, R. D., Bailey, A. S., Campbell, J. L., & Hayward, J. (2020). Quantifying uncertainty in annual runoff due to missing data. *PeerJ*, *8*. https://doi.org/10.7717/peerj.9531

Vieux, B. E., & Moreda, F. G. (2003). Nutrient loading assessment in the Illinois River using a synthetic approach. *Journal of the American Water Resources Association*, *39*(4), 757–769. https://doi.org/10.1111/j.1752-1688.2003.tb04403.x

Vlah, M. J., Rhea, S., Bernhardt, E. S., Slaughter, W., Gubbins, N., DelVecchia, A. G., Thellman, A., & Ross, M. R. V. (2023). MacroSheds: A synthesis of long-term biogeochemical, hydroclimatic, and geospatial data from small watershed ecosystem studies. *Limnology And Oceanography Letters*, *8*(3), 419–452. https://doi.org/10.1002/lol2.10325

Zeileis, A. (n.d.). *zoo : An S3 Class and Methods for Indexed Totally Ordered Observations*.

Zhang, Q., & Hirsch, R. M. (2019). River Water-Quality Concentration and Flux Estimation Can be Improved by Accounting for Serial Correlation Through an Autoregressive Model. *Water Resources Research*, *55*(11), 9705–9723. https://doi.org/10.1029/2019WR025338

Zimmer, M. A., Pellerin, B., Burns, D. A., & Petrochenkov, G. (2019). Temporal variability in nitrate-discharge relationships in large rivers as revealed by high-frequency data. *Water Resources Research*. https://doi.org/10.1029/2018WR023478

6 Supplement

Supplemental Table 1: A table of the results from the synthetic time series experiments. All values (except for number of outliers, which is a count) are expressed in percent error.

(supplemental table 1 is available as supplemental_table_1.csv here 10.6084/m9.figshare.24991455)