A User-Focused Assessment of Annual Load Estimation Methods in Small Watersheds

Gubbins, Wes, Mike, Spencer, Matt, Emily (need Matt feedback on this order)

# Introduction

Annual stream load is the mass of solutes or sediments that move past a point in a stream during a water year. Load is the product of the solute’s concentration and streamflow, integrated over time, as shown in equation 1.

Where is the load, the concentration, the streamflow, and the time. Quantifying how and when solutes move out of watersheds gives researchers key insights into how a watershed functions as a system. Quality solute load estimates are used throughout the field of watershed science. Researchers interested in nutrients rely on estimates of solute export to compare watershed function 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, 2104). Land managers and government decision makers rely on accurate estimation of solute loads to write and apply policy (USEPA, 2000; Schilling et al, 2017).

Through over a century of collective effort, streamflow gauging and modelling can now generate reliable, near-continuous estimates of discharge for most systems of interest. Measuring, or modeling, the concentrations of the myriad solutes of interest to water researchers near continuously has proven to be a more persistent challenge. As most solutes require their own laboratory test to measure, chemistry samples have traditionally been taken as discrete, bottle samples. As these samples need to be manually collected and analyzed, sampling frequencies have been on the order of weekly or monthly for many watershed ecosystem studies.

Converting discrete measurements of stream chemistry and near-continuous discharge into a truly continuous estimate of load introduces uncertainty into those estimates. Many different methods have been used to compute these estimates, ranging from simple averages to complex statistical models. Previous work has shown that estimating loads accurately is complicated by inconsistent sampling intervals (Schilling et al, 2017), conflicting methods (Appling et al, 2015), and simply too little underlying chemistry data (Kirchner et al, 2004). All of these complications arise from the uncertainty in modeling solute concentrations between measurements.

In recent decades, the technology to monitor water chemistry via high-frequency sondes have become cost-effective enough (for a limited range of solutes) to be measured at the same frequencies as discharge is modeled. However, for researchers interested in pre-sensor time series, there is currently no reliable substitute for observations collected in-situ. This puts water researchers, especially those interested in synthesis science, into a quandary. When are load estimates made under a discrete sampling regime comparable to loads measured with more modern methods? The complications mentioned previously are further convoluted when applied across many different sites and sampling procedures. These issues are especially vexing for synthesis scientists, as they are often unable to rely on the intimate, working site knowledge that informs single site analyses.

Many efforts have been taken to improve the accuracy of solute load estimates (Appling et al 2015; Aulenbach et al, 2016; Shilling et al, 2017), but those efforts have either focused large watersheds (Appling et al, 2016; Schilling et al, 2017) or on choosing the best method possible for a given time series (Aulenbach et al, 2016). While both are deeply important for the field at large, they do not clearly delineate when load estimates can be used in cross-site comparisons. This study seeks to build on these prior efforts with three goals. (1) To provide clear guidance to synthesis scientists on what degree of confidence to have in load estimates originating from small watershed studies, (2) to provide flux estimates for a synthesis dataset of small watershed studies and, (3) provide a framework for classifying potentially comparable flux estimates for synthesis science.

# Methods

Our study first uses a case study of data from the Hubbard Brook Experimental Forest (HBEF) to show the relative effect of methods on load estimation error. We then performed a statistical experiment on idealized, synthetic time series data to illustrate the ranges uncertainty possible from an array of methods commonly used to generate load estimates from non-sensor records. Lastly, we apply our tested methods to the geographically diverse, Macrosheds synthesis dataset (Vlah et al, in review) to explore the relative effect these methods have on conclusions made in synthesis science.

## Load Estimation

Four common load methods were chosen for this study, linear interpolation (LI), Beale ratio, rating, and composite. These methods were chosen as they are archetypical of the array methods used commonly 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).

### Linear Interpolation

Linear interpolation was performed by linearly interpolating sampled chemistry values to match their accompanying discharge time series, then computing loads by summing the time series, as represented in equation 2.

Where is the load in kg/ha/year, is the interpolated concentration in mg/L, and is the streamflow in Lps. Due to its simplicity, linear interpolation is commonly used in studies, especially where load estimation methods are not the focus (Kahl et al, 1999; Coombs and Melack, 2012). We used the ‘RiverLoad’ R package’s ‘method6’ function to apply linear interpolation.

### Beale Ratio Estimator

The Beale ratio estimator is often chosen for use in studies for its ability to provide unbiased results (Nava et al, 2019). This method first calculates a mean daily load with a product of concentration and discharge for days with data for both parameters. Then, a ratio of total flows sampled for chemistry over total flows observed and information on how and covary is used to scale the resulting load estimate, as represented in equation 3.

Where is the mean flow for when concentration was measured and is the mean 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.

### Rating

The rating method first relates concentration to discharge in log-log space with a simple linear model. Then the resulting least square 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 concentration-discharge relationship (Crawford, 1999; Quilbe et al, 2006). We used the ‘RiverLoad’ R package’s ‘rating’ function to generate rating estimates.

### Composite

The composite method follows the same method 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 flow at that time are calculated. The residuals are then applied as a correction to the time series at each time step. 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.

## Data Coarsening

To estimate each method’s sensitivity to sampling frequency, we performed a data coarsening experiments on two time series from the watershed 3 site at HBEF, one of nitrate and one of calcium. Both 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 calcium time series was constructed by fitting a least square regression line with no intercept between specific conductance sensor readings and calcium grab samples taken at the site. The discharge time series had a mean flow of 8.94 Lps, a flow standard deviation of 19.20 Lps, and 2.6 x 1010 liters of yield for the year. A plot of flow for the water year is available in supplementary figure 1. The nitrate time series had a mean concentration of 0.048 mg/L with a standard deviation of 0.032 mg/L. The calcium time series had a mean concentration of 0.86 mg/L with a standard deviation of 0.23 mg/L. Both chemistry samples had at least 75 observations on every day, with all days except for 10 having a full 96 samples per day. A plot of both chemistry time series is available in supplementary figure 2.

### Coarsening Procedure

We coarsened each time series from the full resolution to daily by hour, from daily to weekly by day, and then to monthly and bimonthly discretely. A random starting point from within each first coarsening interval was chosen and every nth sample was then taken to create a coarsened time series. We then applied each method to each coarsened time series to generate annual flux estimates. This process was repeated 100 times to generate an envelope of possible estimates for each method.

### 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.

## Synthetic Time Series

Due to the limited availability of high-quality, high-frequency senor and streamflow data from small watersheds, synthetic time series were created to test the sensitives of each method to various hydrologic and concentration-discharge relationships. Past work has shown that the best available load estimation method depends largely on data density and solute chemodynamics (Aulenbach et al, 2016).

Knowing the chemodynamics of the solute of interest is critical to generating a quality load estimate (Aulenbach et al, 2016). In large, interconnected watersheds, many solutes are chemostatic (Godsey et al, 2019). This is less common in small watersheds. In chemostasis, variation in solute concentration is low and does not vary with streamflow. Solutes can also display no pattern with streamflow, but vary widely due to other factors. This is often true of nutrients, such as nitrate. Other solutes, like dissolved organic matter, often increase with increasing discharge, or are called enriching. The opposite is true of many geochemical solutes, such as magnesium or potassium. Instead, they often dilute as flows increase (Godsey et al, 2019).

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 concentration-discharge 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 FINAL GITHUB LINK.

### 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. Then, residuals from the model were randomly reshuffled and applied to the original time series, as described in equation 4.

Where is the resulting for the day, is the observed discharge at that time, is the randomly reshuffled residual for that day, *k* is a constant to prevent zero flow days, and the ratio term is a hold factor to set the total water yield for the year the same for each generated time series.

The same method was then applied to generate stormflow dominated and baseflow dominated time series, as described in equations 5 and 6 respectively.

Where the exponent applied to 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.

We used the ‘EcoHydRology’ R package’s ‘BaseflowSeperation’ function to determine proportion quickflow to total flow for each time series. The unaltered time series had a ~20% quickflow, the stormflow time series had ~35% quickflow, and the baseflow time series had ~5% quickflow.

### Concentration-Discharge Regimes

To generate time series of chemostatic 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 timeseries. 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 timeseries.

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

Similarly, we applied equation 8 to generate time series of dilution.

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

### Coarsening and Load Calculations

First, we calculated ‘true’ flux for each synthetic site-year using the full synthetic time series and the composite method (as recommended in Aulenbach et al 2016) to estimate over less than complete 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.

## Application to Macrosheds Dataset

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 macrosheds.org. The dataset includes harmonized data from 169 federally funded watershed studies from across the United States (Vlah et al, in review). To provide flux estimates to the broader community, we applied all four 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. 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.

# Results

## Case Study of Method Sensitivity to Sampling Frequency

Results from the data coarsening experiments are represented in figures 1 and 2. Generally, all methods struggled to estimate nitrate load. Nitrate and discharge did not have a pattern through the 2016 water year, with the least square regression line in log-log space having a slope of 0.11 and an r-squared of 0.07.

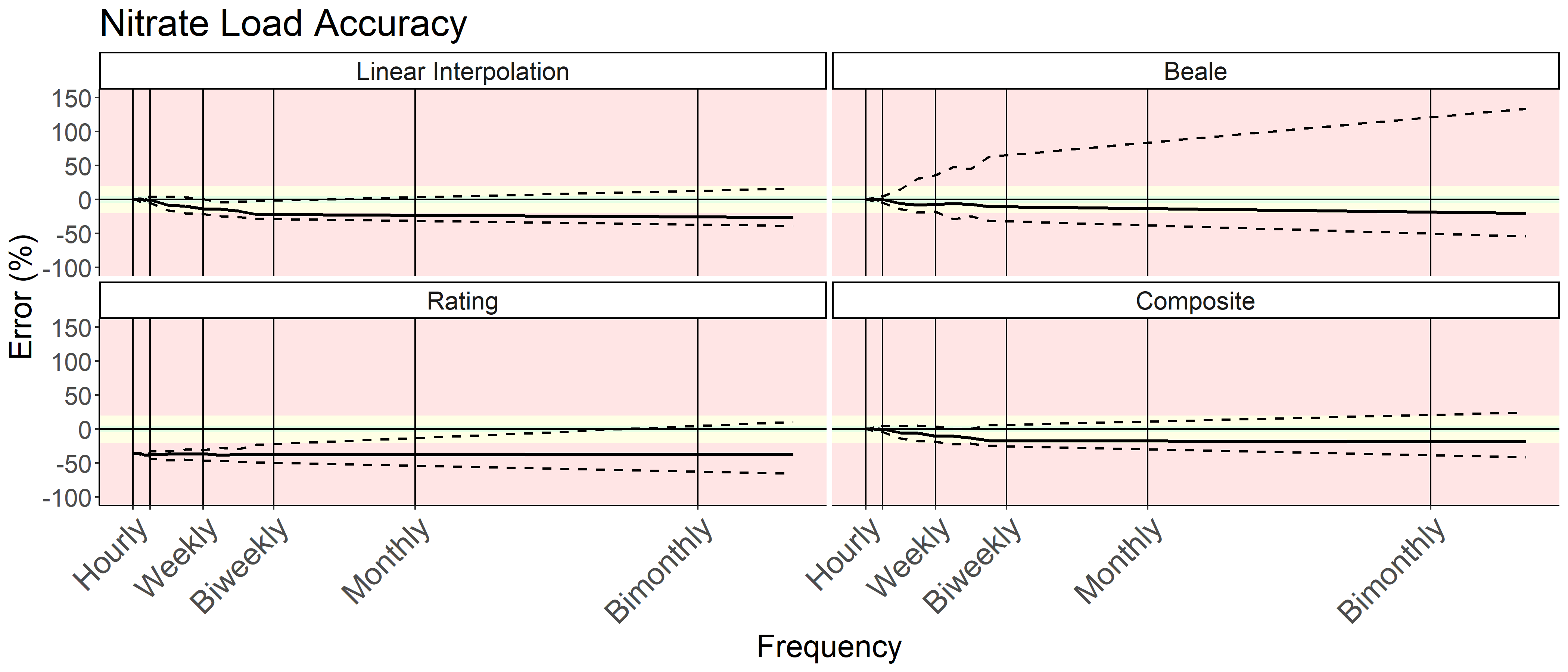


Figure 1: Results from applying four different load estimation methods to artificially coarsened, high-frequency sensor time series of nitrate. Percent error of estimated annual load from the ‘true’ load, calculated using the full time series, is on the y-axis. Coarsened data frequency is on the x-axis. The solid line indicates the median error for that frequency, with dashed lines indicating minimum and maximum. Error within 5% of truth is shaded in green, with 20% yellow, and beyond 20% red. Note that the linear interpolation and composite methods perform relatively the best for this year of data. However, all methods performed poorly with less than daily data.

There was a tight relationship between calcium 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. Calcium and discharge had also had a strong relationship, with the least square regression line in log-log space having a slope of -0.12 and an r-squared of 0.79. The error resulting from each method over each coarsening interval is shown in figure 2.

## 

Figure 2: Results from applying four different load estimation methods to artificially coarsened, high-frequency sensor time series of calcium. 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 dashed lines indicating minimum and maximum. Error within 5% of truth is shaded in green, with 20% yellow, and beyond 20% red. Note that the 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.

In both the nitrate and the calcium time series, the C:Q relationship is not fully informative. The rating methods in both experiments chronically underpredict true load. Inspecting the C:Q relationships in supplementary figures 3 and 4 elucidates this effect. In both cases, the C:Q plots show a minor enriching trend at high discharges that is not captured in the simple linear models developed.

## Synthetic Time Series

The results for our synthetic time series experiments are shown in figure 3. Generally, no pattern data generated the least accurate estimates, regardless of method. All methods degraded with increasingly coarsened data. A full table of the results from figure 3 are available in supplementary table 1.

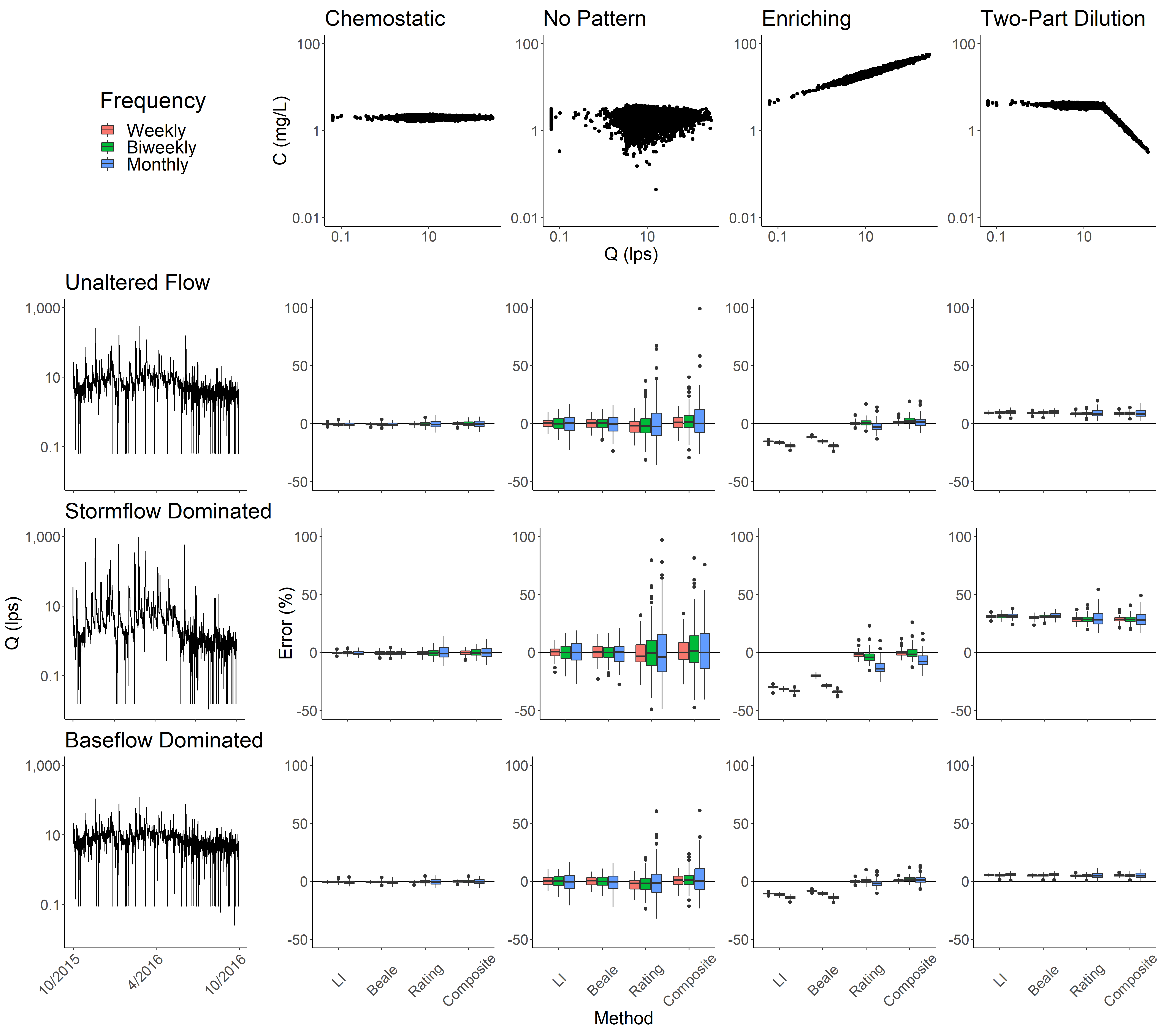


Figure 3: Results from the synthetic time series experiments. The top row of plots describes the concentration discharge relationship. The side column of plots describes the hydrologic regime. The grid boxplots is the result of applying the four methods across each combination of concentration discharge relationship and hydrologic regime. Note that no-pattern consistently shows high error, while chemostatic, diluting (with composite method and high data frequency), and enriching (with composite method and high data frequency) give low error.

## Application to Macrosheds Dataset

Applying our methods to the Macrosheds dataset generated 16,489 site-years of data across 93 sites and 112 solutes. The load calculations from each site-year of Macrosheds data are available at FIGSHARE LINK, and the latest dataset is available at macrosheds.org.

# Discussion

## Validation of Past Efforts and Importance of Concentration-Discharge Relationships

Our results generally confirm what Aulenbach and others (2016) observed in their assessment of load estimation methods in small watersheds. Namely, that when there is not a strong C:Q relationship, users should rely on linear interpolation or averaging methods. This is evident in figure 3 under chemostatic or no-pattern C:Q relationships. Under these C:Q relationships, linear interpolation and the Beale estimator outperform the rating and composite methods. When there is a strong C:Q relationship, users should use a method that leverages that relationship. Our results in figure 3 confirm that when the C:Q relationship is stable and effectively modeled, C:Q informed methods outperform others. Under such conditions, the rating and composite methods dramatically outperform linear interpolation or the Beale ratio estimator. However, it should be noted that these methods did not apply as neatly to the actual, more complex C:Q relationship present in the coarsened calcium time series (figure 2). As shown in figure 3, more flashy hydrologic regimes increase variance across all methods.

This study also validates the call made by Kirchner and others in 2004 to continue to prioritize high-frequency measurements of water quality. Regardless of estimation method, data of higher frequency and quality greatly improves load estimation uncertainty.

## Challenges for Synthesis Science with Load Estimates

The results of our efforts highlight serval important challenges for synthesis scientists working with load estimates.

The first, is that method selection is deeply important for accurate load estimation. The results in figure 2 clearly show that, especially at coarse sampling intervals, method selection greatly influence the potential range of estimates generated. Looking at two solute case studies from the Macrosheds dataset in figure 4, we can see that choosing the correct method can greatly sway the conclusions made from a load estimate. Both figures highlight the absolute need to know the underlying method was used to calculate a load estimate before including it in subsequent analyses.

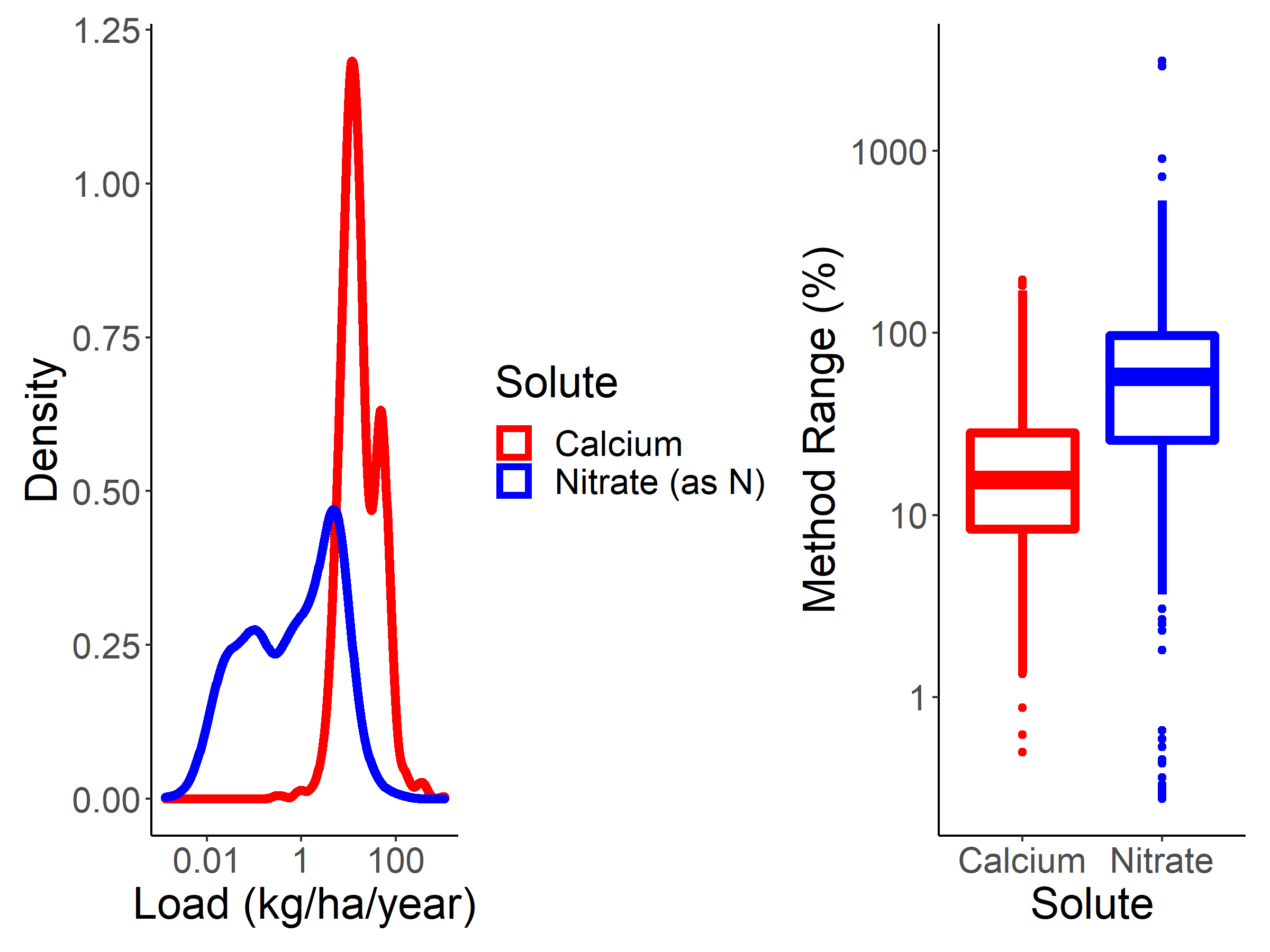


Figure 4: A case study of how method selection varies by solute in the Macrosheds dataset. On the left is a density diagram of annual load estimates made using the Aulenbach et al, 2016 selected method. On the right is a boxplot of the total range the methods in this study produced over the Macrosheds site-years. Method range was computed by taking the percent difference of the maximum and minimum estimates for each site-year. Calcium used 1,065 site-years of data and had a median method range of 15.6%. Nitrate used 1,123 site-years and had a median method range of 57.4%.

When a load estimate is provided without documentation of the underlying methodology, it should not be trusted. For example, extrapolating from the results in figure 4, a researcher working with an unknown method on a nitrate load estimate from a small watershed should reasonably assume baseline uncertainty between 25.8% and 97.2% (the second and fourth method range quartiles respectively). Synthesis scientists are better served by using the underlying chemistry and discharge time series to recompute load undocumented load estimates using known methods. The work necessary to do so can be eased using tools like the ‘Riverload’ and ‘Macrosheds’ R packages.

Additionally, our experiments give substantial evidence that at sampling frequencies commonly used in ecosystem studies (in the range of weekly to bimonthly) load estimates should be used with careful consideration, regardless of the site hydrology or C:Q relationship (as seen in figure 3). In our synthetic time series analysis, we regularly observed >10% error values even with the most appropriate methods and near-ideal C:Q relationships. While the appropriate method at a weekly sampling frequency would often produce a low mean error, the extent of possible error observed over 100 runs was still problematic. For example, under baseflow dominated hydrologic regime, with a chemostatic C:Q relationship, and weekly sampling, linear interpolation (the most appropriate method) gave a mean error of -1.2%, but a maximum error of 10.2% and a minimum error of -12%. The second and third error quantiles are at -8.6% and 4.3%, respectively.

Finally, knowing fit of the C:Q relationship is not enough to assign confidence their underlying load estimates. Users need to first assess how high their confidence is in knowing the entirety of the C:Q relationship. For example, if a user were to assess the strength of the C:Q relationship of the two-part dilution time series with weekly sampling in a stormflow dominated system, users would find a strong fit (as defined by Aulenbach et al 2016). Without knowing the full shape of the relationship, they would be inclined to leverage that strong fit via the composite method. However, this would give the user artificially high confidence in their estimate, and they should instead get more accurate results with linear interpolation. Similarly, a site that biases its collection towards baseflow days (which is common with non-event supplemented sampling) may erroneously conclude that their C:Q relationship is chemostatic or has no pattern. A user making this conclusion would be tempted to use linear interpolation or the Beale ratio estimator to reduce both error and bias. This would lead to load estimates that heavily underestimate true load for the year. Put simply, to truly asses your confidence in a load estimate you need to know how confident you are in both the C:Q relationship itself and your 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. 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 observations.

## A Framework for Considering Load Estimates

The challenges from the previous section are easy to point at, but difficult to solve, especially for synthesis scientists. While other researchers can rely on an intimate understanding of the history and disturbances at each site to inform their confidence in estimates, synthesis scientists must rely only on the data output from previous studies. Important site history narratives and key assumptions may be buried deep in a chain of building publications, if 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 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 laid out in figure 5. As C:Q relationships have been shown to change over time (Kirchner et al, 2004) 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.

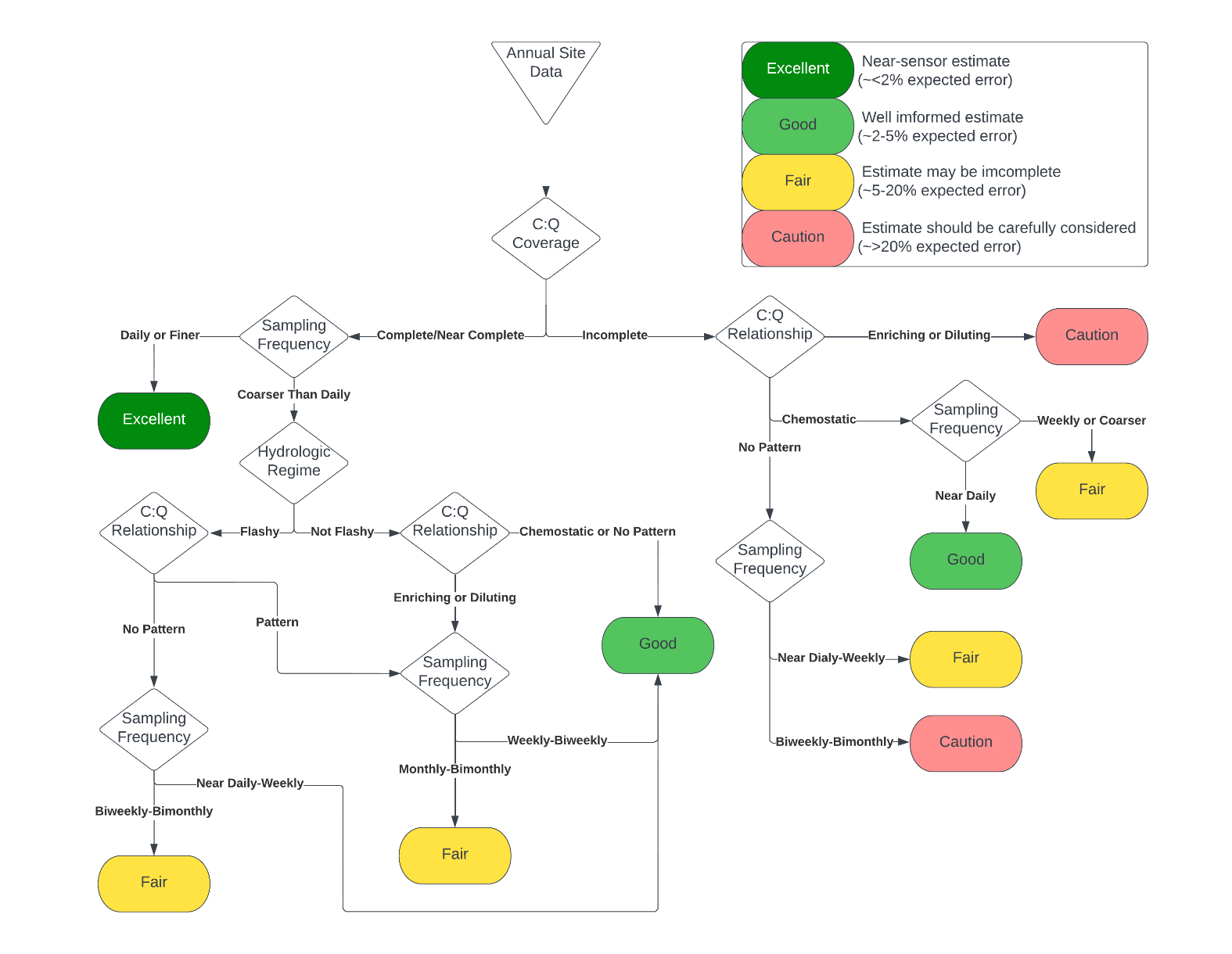


Figure 5: A conceptual flowchart for classifying confidence in load estimates, assuming a complete discharge time series and method selection as described in Aulenbach et al 2016. C:Q coverage refers how completely the extent of the C:Q relationship has been confirmed for the given site-year. Sampling frequency is the rate 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 produce ‘excellent’ estimates. Non-chemostatic site-years with incomplete C:Q coverage produce ‘fair’ estimates or estimates that should be treated with ‘caution’.

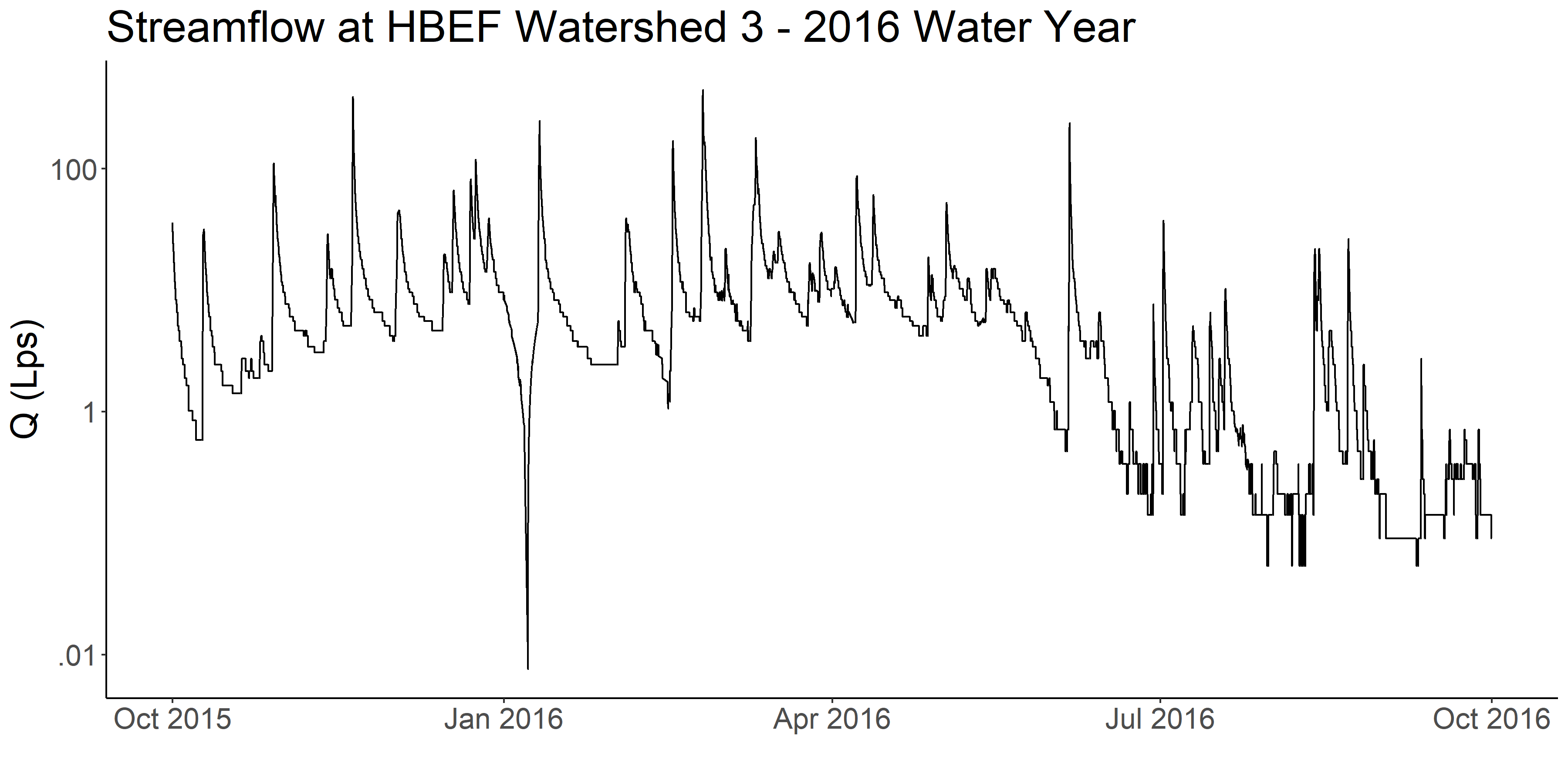
Data binned as ‘excellent’ or ‘good’ is likely suitable for inter-site comparisons at large scales. Data binned as ‘fair’ should only be used for limited applications. For example, ‘fair’ rated estimates could be used in aggregated regional estimates of weathering rates. Data binned as ‘caution’ should not be used without the user learning more about the site. It is possible that ‘caution’ 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 5, we fully expect to be able to do so robustly as more sensor data becomes available in a variety of small watershed systems. In the meantime, we encourage the watershed synthesis science community to begin to use our framework to consider how and when they make inter-site comparisons of solute export.

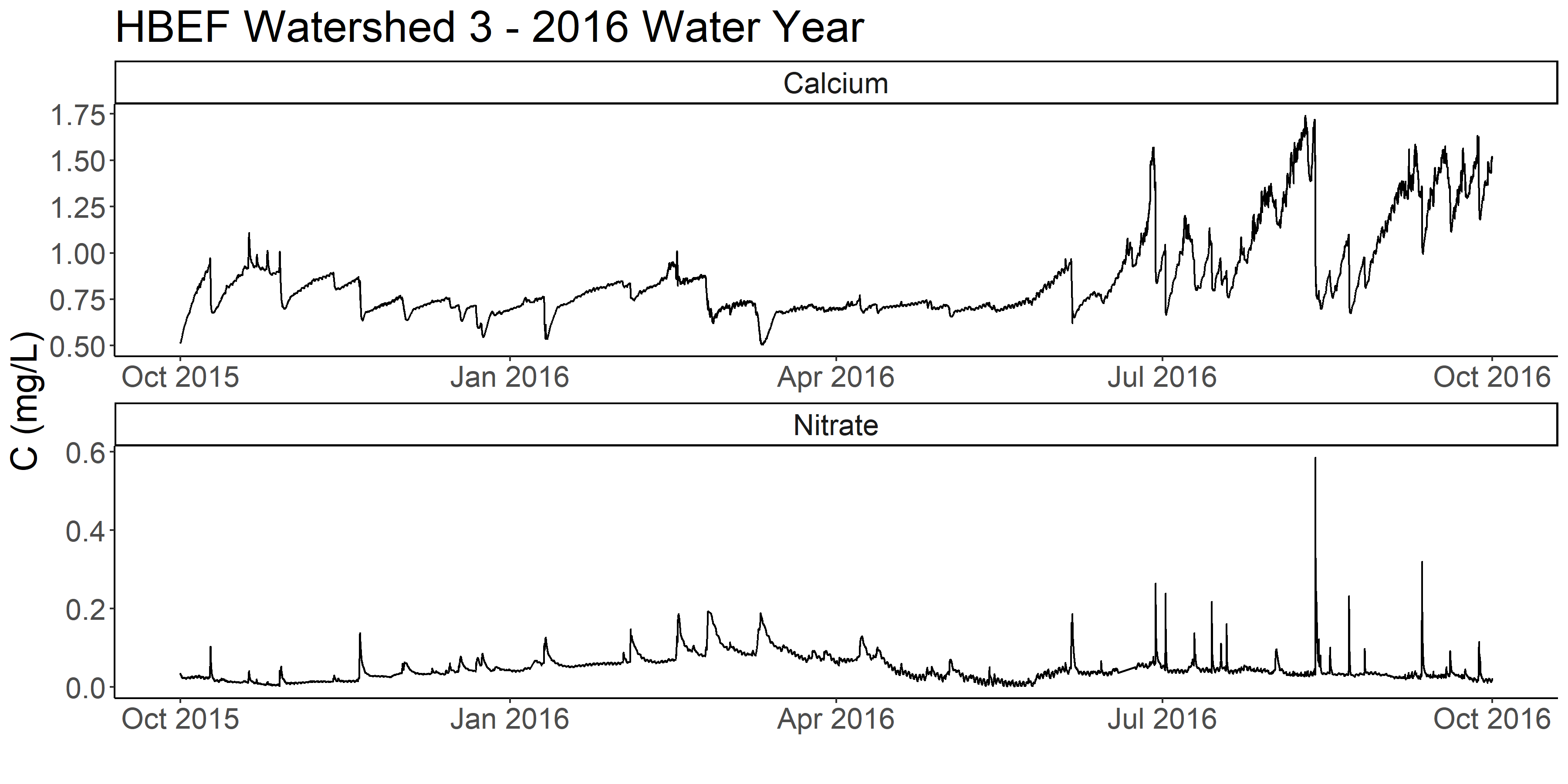
Our framework first asks users to consider the degree of confidence they have in observing the entire C:Q relationship. In its simplest form, this could be accomplished by looking at a site hydrograph with sampling dates marked on it. Systematically, this could be assessed by creating a ratio of sampled 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 chemistry time series standard deviation to mean. Hydrologic regime can be assessed using the Richards-Baker flashiness index (Baker et al, 2004).

We hope the framework provided here can unlock synthesis 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.

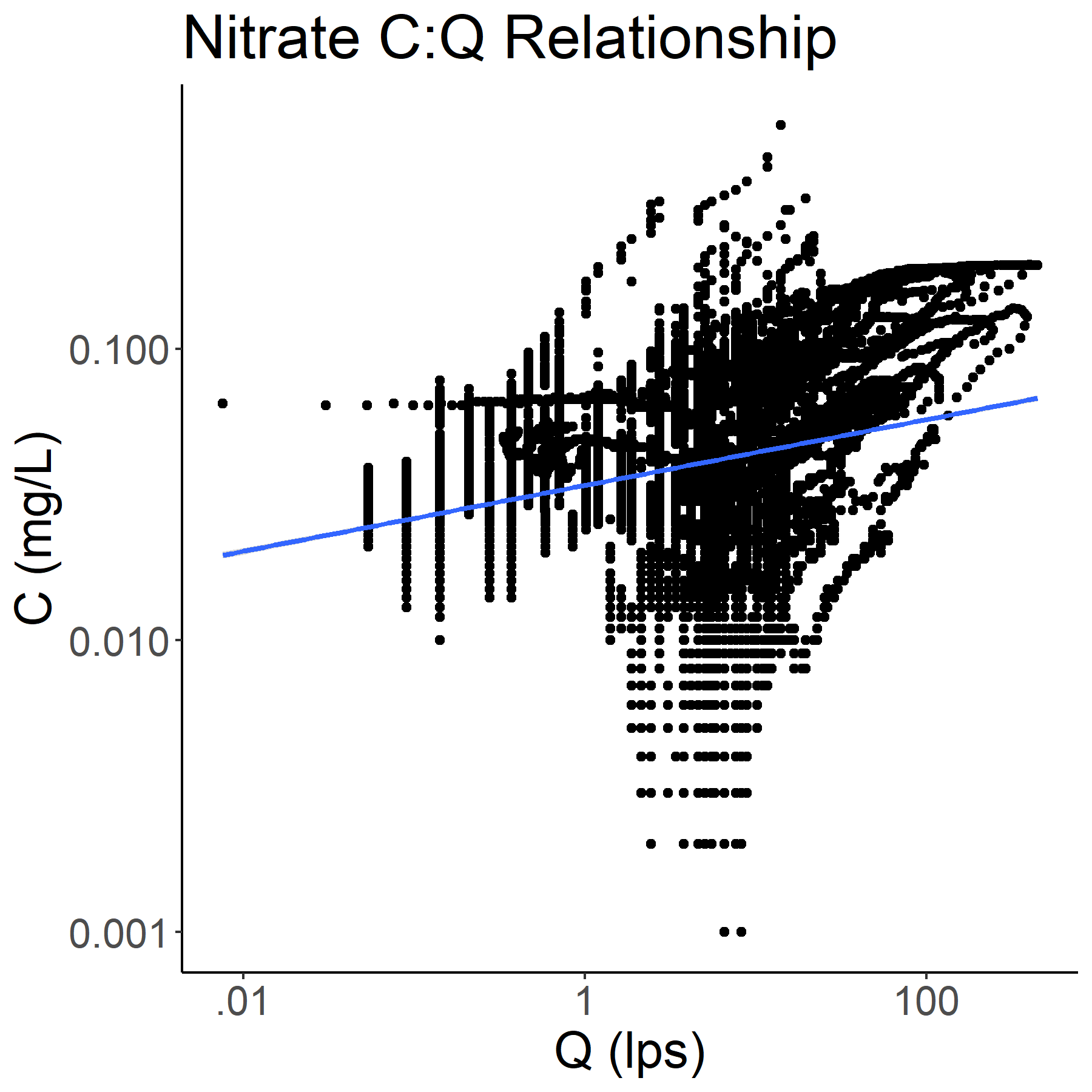
# Supplement



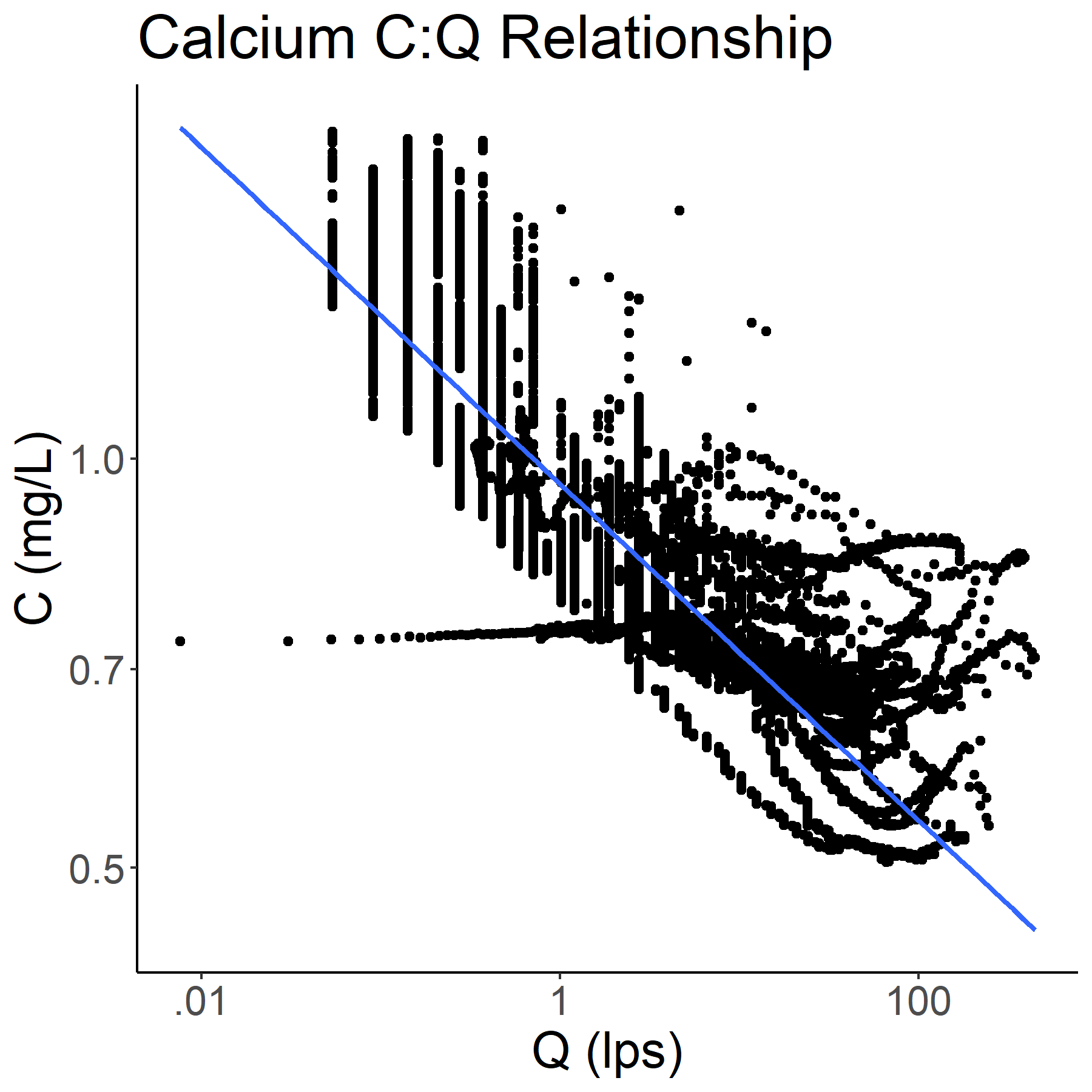
Supplemental Figure 1: The underlying streamflow data used in the data coarsening and synthetic time series experiments. Data was collected at Hubbard Brook Experimental Forest in watershed 3 using a long-running rating, v-notch weir, and pressure transducer.



Supplemental Figure 2: The underlying chemistry time series used in the data coarsening experiments. Data was collected at Hubbard Brook Experimental Forest in watershed 3 using a multiparameter sonde. The record has no missing days.



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



Supplemental Figure 4: Concentration-discharge plot for calcium at HBEF watershed 3 for the 2016 water year. Observations are in black and the least squares regression line used in the rating and composite methods is in blue. The slope of the best fit line is -0.12 with an r-squared of 0.79. The calcium time series has a mean of 0.86 mg/L and a standard deviation of 0.22 mg/L.