12-15-2015

Reducing bias and quantifying uncertainty in watershed flux estimates: the R package loadflex

Alison P. Appling
University of New Hampshire, Durham

Miguel C. Leon
University of Pennsylvania

William H. McDowell
University of New Hampshire, Durham, bill.mcdowell@unh.edu

Recommended Citation

Appling, A. P., M. C. Leon, and W. H. McDowell. 2015. Reducing bias and quantifying uncertainty in watershed flux estimates: the R package loadflex. Ecosphere 6(12):269. https://dx.doi.org/10.1890/ES14-00517.1

esa

ECOSPHERE

SPECIAL FEATURE: UNCERTAINTY ANALYSIS

Reducing bias and quantifying uncertainty in watershed flux estimates: the R package loadflex

ALISON P. APPLING,[1,3,†] MIGUEL C. LEON,[2] AND WILLIAM H. MCDOWELL[1]

[1]*Department of Natural Resources and the Environment, University of New Hampshire, Durham, New Hampshire 03824 USA*
[2]*Department of Earth and Environmental Science, University of Pennsylvania, Philadelphia, Pennsylvania 19104 USA*

Citation: Appling, A. P., M. C. Leon, and W. H. McDowell. 2015. Reducing bias and quantifying uncertainty in watershed flux estimates: the R package loadflex. Ecosphere 6(12):269. http://dx.doi.org/10.1890/ES14-00517.1

Abstract. Many ecological insights into the function of rivers and watersheds emerge from quantifying the flux of solutes or suspended materials in rivers. Numerous methods for flux estimation have been described, and each has its strengths and weaknesses. Currently, the largest practical challenges in flux estimation are to select among these methods and to implement or apply whichever method is chosen. To ease this process of method selection and application, we have written an R software package called `loadflex` that implements several of the most popular methods for flux estimation, including regressions, interpolations, and the special case of interpolation known as the period-weighted approach. Our package also implements a lesser-known and empirically promising approach called the "composite method," to which we have added an algorithm for estimating prediction uncertainty. Here we describe the structure and key features of `loadflex`, with a special emphasis on the rationale and details of our composite method implementation. We then demonstrate the use of `loadflex` by fitting four different models to nitrate data from the Lamprey River in southeastern New Hampshire, where two large floods in 2006–2007 are hypothesized to have driven a long-term shift in nitrate concentrations and fluxes from the watershed. The models each give believable estimates, and yet they yield different answers for whether and how the floods altered nitrate loads. In general, the best modeling approach for each new dataset will depend on the specific site and solute of interest, and researchers need to make an informed choice among the many possible models. Our package addresses this need by making it simple to apply and compare multiple load estimation models, ultimately allowing researchers to estimate riverine concentrations and fluxes with greater ease and accuracy.

Key words: composite method; concentration; constituent; flux; nutrient; R; software; solute load models; uncertainty; watershed.

Received 18 December 2014; revised 26 May 2015; accepted 4 June 2015; **published** 15 December 2015. Corresponding Editor: J. Taylor.

[3] Present address: Center for Limnology, University of Wisconsin, Madison, Wisconsin 53706 USA.

† **E-mail:** alison.appling@gmail.com

INTRODUCTION

Quantifying solute and sediment fluxes from watersheds can yield insights into watershed processes, in-stream processes, and nutrient cycles at regional to global scales. Flux estimates are an essential component of whole-watershed manipulation experiments (Likens et al. 1970), measurements of nutrient retention within catchments (Groffman et al. 2004) and within stream

reaches (Bowes and House 2001), calibration of global nutrient export models (McCrackin et al. 2014), and assessments of export trends across time and space (Hruška et al. 2009, Prokushkin et al. 2011). A wide variety of quantification methods have been proposed, including simple period-weighted means, ratio estimators, binning methods, and both linear and nonlinear statistical models (Preston et al. 1989, Letcher et al. 1999, Asselman 2000, Cox et al. 2008, Birgand et al. 2010, Raymond and Saiers 2010, Worrall et al. 2013). Complex methods have been made more accessible through implementation software packages including FLUX (Walker 1996), LOADEST (Runkel et al. 2004), and EGRET (Hirsch et al. 2010, Hirsch and De Cicco 2015). These approaches often involve some sort of regression model, for which a primary challenge is to identify an adequate model formula to represent the relationship between fluxes (or concentrations) and available predictors such as discharge, season, or year. While regression models aid our conceptual understanding of the controls on watershed solute export, they can generate poor flux estimates when the few available predictors fail to capture all causes of variability in flux. Model inadequacies often take the form of short-term biases, in which predictions for a certain period are consistently below or consistently above the actual values (Hirsch 2014).

One solution for better estimating watershed fluxes and concentrations, commonly termed the "composite method," has been recently introduced (Huntington et al. 1994), named and described (Aulenbach and Hooper 2006), applied (Peters et al. 2006), extended (Verma et al. 2012), and explored (Aulenbach et al. 2007, Aulenbach 2013). The composite method combines the predictions from a regression model with an empirical "residuals correction" to bring predictions closer to observations during the period of interest. This two-step process can reduce short-term biases and thereby lead to more accurate estimates of total fluxes or mean concentrations (Aulenbach and Hooper 2006, Verma et al. 2012, Aulenbach 2013). The resulting estimates can then be used to evaluate the contribution of a watershed to total regional or global solute loads, to more accurately describe the pattern of concentrations or fluxes over time, or to make inferences about historical fluxes from sparse data records. Despite the potential for the composite method to improve flux and concentration estimates, however, the method has remained out of reach to many scientists because existing implementations of the method are private or implemented with proprietary software.

Here we present the first public, open source implementation of the composite method, complete with diagnostic tools to aid researchers in applying the composite method to estimate watershed fluxes. Our package, titled `loadflex` and implemented in the R statistical language, is freely available at https://github.com/mcdowelllab/loadflex. In fact, the `loadflex` package implements more than just the composite method: the package also contains methods for interpolating among observations (these methods may also be called integration methods or period-weighted approaches), methods for applying linear regressions of arbitrary complexity, and an interface to the suite of models made available by the USGS as the `LOADEST` package in Fortran (Runkel et al. 2004) or, more recently, the R implementation named `rloadest` (Lorenz et al. 2015).

A key strength of `loadflex` is that it implements each of these load models with a common interface, enabling users to quickly fit and compare several models. This is useful because the best approach for each new river and dataset depends on the hydrology of the watershed, the sources and mobility of the solute of interest, the data resolution and precision, and the degree of autocorrelation in the residuals of proposed models (Letcher et al. 1999, Johnes 2007, Aulenbach 2013). Another paper in review for this special issue of Ecosphere proposes general guidelines for choosing among the composite method, a period-weighted approach, and a regression model (Aulenbach, *unpublished manuscript*). A more site-specific approach is to fit several models to the data and compare them for goodness of fit, theoretical utility, and realism of predictions. However, the time and energy required to learn and implement each new model can prevent researchers from thoroughly exploring their modeling options. We designed `loadflex` to empower researchers to explore more competing models, to make informed decisions among more options, and, ultimately, to make

better flux and concentration predictions.

A second strength of the `loadflex` package is that it computes not just central estimates of solute fluxes but also their uncertainty, which is essential to interpreting flux estimates (Yanai et al. 2015). For the composite method, we introduce a new method for estimating the uncertainty in these estimates. Previous studies have repeatedly fitted composite models to subsets of the data to argue that the composite method should have lower uncertainty than a regression model alone; however, no independent estimates of uncertainty have been hitherto available (Aulenbach and Hooper 2006, Verma et al. 2012, Aulenbach 2013). Whereas prediction uncertainty for regression models can be estimated by standard functions of the fitted regression parameters and residuals, such methods are inappropriate for composite models, for which the estimates are substantially altered from the original regression predictions. Even common non-parametric methods such as a simple jackknife or bootstrap are inappropriate because of the time-dependence of the composite method correction. One solution, which we have implemented and will describe here, is to combine a parametric bootstrap with a delete-one jackknife to estimate the overall standard error of prediction. This solution represents a new addition to the composite method literature.

Here we introduce and demonstrate our `loadflex` package. We describe the implementation of four general classes of load models and their associated methods for assessment and prediction. We emphasize our enhancements to the composite method, which include additional options for the residuals correction, compatibility with models from the USGS `rloadest` package, and our new algorithm to estimate uncertainty. To demonstrate the package, we fitted four models to 12 years of weekly data from the Lamprey River in New Hampshire, finding that the choice of model can substantially affect conclusions about the effects of floods on nitrate fluxes in this particular watershed. We then use two years of high-resolution sensor data to more rigorously evaluate the four models and their uncertainty estimates. The `loadflex` package makes it possible to apply and evaluate several common and promising models, to choose the best model for the data at hand, and to make predictions with a strong understanding of their accuracy and uncertainty.

Methods and Demonstration

Overview

`loadflex` provides a standardized workflow for fitting and applying a model of solute concentrations or fluxes, regardless of the specific model that is used. The user (1) fits a regression or interpolation model using a function named for the chosen model; (2) assesses the model and data using standard techniques in R; (3) if appropriate, applies the composite method to the original model with one additional command; (4) uses the final model to make point predictions about concentrations or fluxes over time, and (5) optionally aggregates the point predictions to longer periods such as months, seasons, or years. Here we demonstrate these steps for four different load models in parallel; essential code is given in Fig. 1, and the complete code files are available as a supplement.

Site description

The Lamprey River is an 81-km river flowing through southeastern New Hampshire. Its 548-km^2 watershed empties into the Great Bay estuary, whose nitrogen impairment has been the subject of much interest and management in recent years (Trowbridge et al. 2014). In its entirety, the watershed is 71.7% forest, 8.6% wetland, and 7.9% agricultural land (NOAA 2006), with 6.2% impervious surface cover (NH GRANIT 2011) and a population density of 67 people per square kilometer as of 2010 (U.S. Census Bureau 2011). The mean annual temperature is 8.2°C and mean annual precipitation is 1206 mm/yr (National Climatic Data Center normals for 1981–2010 at Durham [USW00054795] and Epping [USC00272800], NH; http://www.ncdc.noaa.gov/cdo-web/datatools/normals).

The nitrate concentrations and discharge data for this study (Fig. 2) were collected at two neighboring sites along the Lamprey mainstem. The Packers Falls site (43°06′09″ N, 70°57′11″ W; NAD27) is 7.5 km upstream of the Great Bay, has a catchment area of 479 km^2, and had mean annual discharge of 8.85 m^3/s in 1981–2010. The Wiswall Dam site is 1.3 km upstream from Packers Falls, has a catchment area of 477 km^2,

```
# A) Create a metadata description of the dataset & desired output
meta <- metadata(constituent="NO3", flow="DISCHARGE",
  dates="DATE", conc.units="mg L^-1", flow.units="cfs", load.units="kg",
  load.rate.units="kg d^-1", station="Lamprey River, NH")

# B) Fit four models: interpolation, linear, rloadest, and composite
no3_interp <- loadInterp(interp.format="conc",
  interp.fun=rectangularInterpolation, data=intdat, metadata=meta)
no3_lm     <- loadLm(formula=log(NO3) ~ log(DISCHARGE),
  pred.format="conc", data=regdat, metadata=meta, retrans=exp)
no3_reg2   <- loadReg2(loadReg(NO3 ~ model(9), data=regdat,
  flow="DISCHARGE", dates="DATE", time.step="instantaneous",
  flow.units="cfs", conc.units="mg/L", load.units="kg"))
no3_comp   <- loadComp(reg.model=no3_reg2, interp.format="conc",
  interp.data=intdat)

# C) Generate point predictions from each model
preds_i <- predictSolute(no3_interp, "flux", estdat, se.pred=T, date=T)
preds_l <- predictSolute(no3_lm, "flux", estdat, se.pred=T, date=T)
preds_r <- predictSolute(no3_reg2, "flux", estdat, se.pred=T, date=T)
preds_c <- predictSolute(no3_comp, "flux", estdat, se.pred=T, date=T)

# D) A few ways to inspect the models
summary(getFittedModel(no3_lm))
qplot(x=Date, y=Resid, data=getResiduals(no3_interp, newdata=intdat))
residDurbinWatson(no3_reg2, "conc", newdata=regdat, irreg=T)
residDurbinWatson(no3_reg2, "conc", newdata=intdat, irreg=T)
estimateRho(no3_reg2, "conc", newdata=regdat, irreg=T)$rho
estimateRho(no3_reg2, "conc", newdata=intdat, irreg=T)$rho
getCorrectionFraction(no3_comp, "flux", newdat=estdat)

# E) Aggregate from point predictions to monthly predictions
aggs_i <- aggregateSolute(preds_i, meta, "flux rate", "month")
aggs_l <- aggregateSolute(preds_l, meta, "flux rate", "month")
aggs_r <- aggregateSolute(preds_r, meta, "flux rate", "month")
aggs_c <- aggregateSolute(preds_c, meta, "flux rate", "month")
```

Fig. 1. Essential code to fit four different models, make predictions, evaluate the models, and aggregate to monthly estimates. Functions from the `loadflex` package are in red. The full code files used to generate the figures and results of this manuscript are available as a supplement.

and includes a 5.5 m tall concrete dam that was built in 1911 and equipped with a fish ladder in 2011 (New Hampshire Department of Environmental Services 2013). The outlet controls of Wiswall Dam are rarely manipulated except on the two days per year when the fish ladder is opened and closed, and the dam thus has a relatively minor and infrequent influence on the river flow regime (D. Cedarholm, *personal communication*).

Nitrate concentrations have been monitored with weekly and event-based grab samples at Packers Falls since 10 September 1999 and are measured by ion chromatography with suppressed conductivity detection on a Dionex 1000 ICS (Sunnyvale, California, USA). At the Wiswall Dam site, nitrate concentrations have been monitored with a SUNA nitrate sensor (Satlantic, Halifax, Canada) at 15-minute resolution since 7 September 2012. The SUNA measurements have been calibrated with additional weekly grab samples at the Wiswall Dam site, with a post-calibration correlation coefficient of $r^2 = 0.59$ between laboratory measurements and calibrated SUNA predictions.

Discharge records at Packers Falls extend from 1953 to the present (US Geological Survey, site 1073500, waterdata.usgs.gov). Two exceptional flooding events occurred in 2006 and 2007, with flow exceeding 150 m^3 s^{-1} from 14 May to 17 May 2006 and from 16 April to 19 April 2007, each time meeting the criterion for a 100-year flood (FEMA 2008). In addition, an unusually large storm occurred in October 2005. Qualitative inspection of the data suggests that nitrate concentrations and fluxes decreased from the

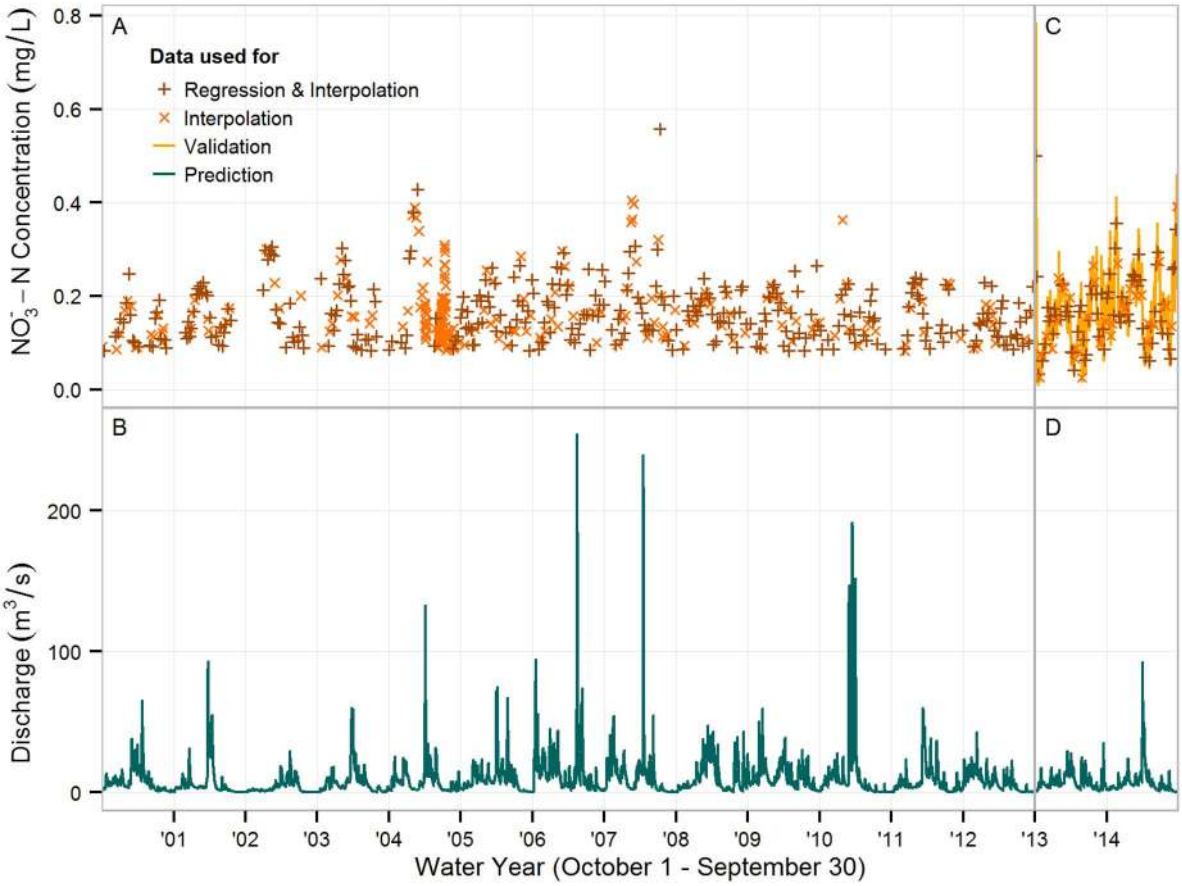

Fig. 2. Nitrate and discharge data from the Lamprey River, New Hampshire, USA. Data in panels A and B are used to demonstrate model fitting, prediction, aggregation, and simple tests of whether nitrate dynamics were altered by the three large floods and storms of 2005–2007. Data in panels C and D are used to compute model performance metrics. Nitrate observations come from grab samples (A) or an in-situ sensor (C). Data points at one-week intervals were selected from the quasi-continuous sensor data to simulate grab samples (× and + in C). Grab samples and simulated grab samples are used to fit the models; each point is used for interpolation only (×) or both interpolation and regression calibration (+), where regression data are a subset of interpolation data to reduce autocorrelation and the consequent underestimation of uncertainty. Discharge observations (B and D) are used both for calibration and prediction and come from USGS gaging station 01073500.

years preceding these events to the years that followed. We hypothesized that the storms and floods could have fundamentally altered flow paths and/or nitrate processing rates in the catchment, possibly by flushing nitrate out of groundwater influenced by septic leach fields or adding woody debris to stream channels as a carbon source for denitrification. Here we ask whether concentrations and fluxes had indeed changed significantly.

Although the Lamprey has sufficient flow to be the type of site most often successfully modeled with regressions on discharge such as those implemented in `LOADEST` or `rloadest`, the best model we could identify within the `rloadest` framework explained only a modest proportion of the overall variance in concentration ($R^2 = 0.38$; see the *`rloadest` regression model* section below). In particular, the Lamprey appears to be an example of a site for which typical regression models are inadequate due to multi-week fluctuations in the discharge-concentration relationship (see *Composite model* below). Due to these multi-week biases in the regression model

predictions, and due to the availability of weekly chemical measurements over the last decade, the Lamprey River is a prime candidate for improved load estimation with the composite method.

Model construction

Interpolation model.—The first and simplest of the four models demonstrated here is the interpolation model. Interpolation models have been used, in various forms, for many studies of solute and sediment fluxes (Porterfield 1972, Grimm 1987, Buso et al. 2000, Vanni et al. 2001). Interpolation models are typically preferred when regression models are unsatisfactory (due to a lack of predictor data or a weak relationship between the available predictor data and the concentrations or fluxes of the solute), especially when direct observations are frequent enough that each pair of successive concentration observations is likely representative of concentrations in the period between them (Webb et al. 1997, Robertson and Roerish 1999).

Here we demonstrate the use of a rectangular interpolation, where horizontal lines are drawn through observations in a plot of concentrations (or fluxes) versus time, and each horizontal line is connected to the next by a vertical line midway between successive observations (example predictions in Fig. 3A). A rectangular interpolation is mathematically equivalent to a period-weighted averaging method (Likens et al. 1977), although the latter method often bypasses the step of making point predictions and moves immediately to estimating the aggregated flux or concentration. The function used to create this model, `loadInterp()` (Fig. 1B), offers the choice of several interpolation methods, including rectangular, piecewise linear (method M6 of Moatar and Meybeck 2005), spline, and smooth spline interpolations. These methods all have in common that they connect observations of concentration or flux over time with a single line, where the values in that line depend only on the specific interpolation function, the observations of concentration or flux, and the date and time.

Custom regression model.—Regression models are a longstanding alternative to interpolation models in estimating watershed solute fluxes (Miller 1951, Johnson et al. 1969, Preston et al. 1989). Whereas interpolation predictions are based entirely on the date/time and the observed concentrations or fluxes, regression models can use a larger set of predictors, such as current and antecedent discharge, season, and year, to make predictions for the dates of interest. Regression models often require less data than interpolation models because the data need only span the range of predictors rather than the full time period of interest (Robertson and Roerish 1999).

In `loadflex`, a linear regression can be fitted using the `loadLm()` command (Fig. 1B and Fig. 3B). Though the regression formula in Fig. 1 contains only a log-discharge term, the formula may contain any number of user-defined variables, including discharge squared, seasonal terms, or time-trend terms. `loadflex` currently requires that the left-hand side of the regression formula (flux or concentration) is logged; this enforces consistency with a later assumption during the prediction phase that the predicted values are lognormally distributed. In future versions of `loadflex` we plan to offer a choice of normal or lognormal distributions for `loadLm` models, e.g., for compatibility with the hyperbolic model of Johnson et al. (1969). However, we expect that the user will often prefer the lognormal assumption, given that the response variable is logged in many rating curves and extrapolation (regression) methods (Miller 1951, Ferguson 1986, Cohn et al. 1989, Johnes 2007).

`rloadest` regression model.—A powerful feature of `loadflex` is the ability to make use of models produced by the USGS package `rloadest` (Lorenz et al. 2015), with a simple wrapper, `loadReg2()`, to get the additional `loadflex` functionality. Relative to simple linear regressions, `rloadest` models are less transparent, but they have several important advantages: they implement several well-established model formulas including the 7-parameter model of Cohn et al. (1992), they are fitted by maximum likelihood methods, and they provide smart handling of censored data (Runkel et al. 2004, Cohn 2005).

Models fitted by `rloadest` are a good option for many watersheds where some sort of regression approach is desired, especially because the weaknesses commonly ascribed to `LOADEST`/`rloadest` models are also present in other log linear regression models, and in many cases have been mitigated in the `LOADEST`/

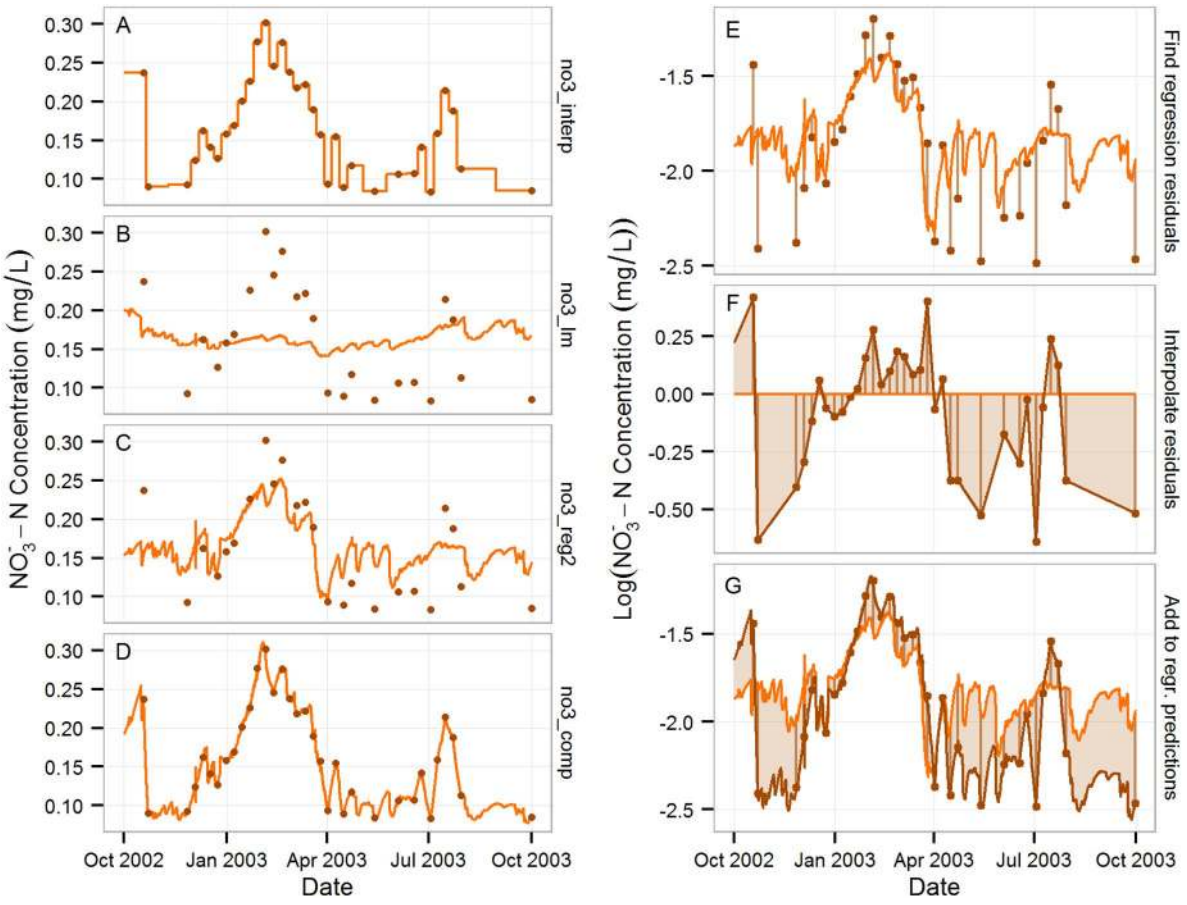

Fig. 3. Models fitted to NO_3^- data from the Lamprey River, New Hampshire, USA, using functions in the `loadflex` package (Fig. 1). Panels A–D show concentration observations (brown points) and concentration predictions (orange lines) for the 2003 water year for (A) an interpolation, (B) a simple linear regression model, (C) an `rloadest` regression model, and (D) a composite model. Panels E–G illustrate the steps taken to fit a composite model, i.e., to move from regression predictions as in C to composite predictions as in D. (E) Regression predictions and observations from C are first log-transformed to compute residuals (vertical lines). (F) Residuals from E are then interpolated, in this case by a piecewise linear function (brown time series line). (G) The composite method predictions (brown line in G) are computed as the sum of the regression predictions (orange lines in E and G) and the interpolated residuals (brown line in F).

`rloadest` implementations. For example, the issue of retransformation bias, discussed in a later section of this manuscript, is automatically addressed in `LOADEST` (Runkel et al. 2004). Another concern is that regression models can be severely biased when model assumptions are violated (Hirsch 2014). These biases arise from non-normality or heteroskedasticity of residuals (Thomas 1988, Stenback et al. 2011) and/or a lack of model fit (Toor et al. 2008, Moyer et al. 2012). When a poor model fit is due to sustained deviations of predictions from true values for periods of weeks to months, that regression model is a good candidate for inclusion in a composite model (next section). Regardless of whether a composite correction is applied, however, model biases can and should be diagnosed in every application of a regression model, and both `LOADEST` and `rloadest` generate outputs to facilitate this process (Runkel 2013, Lorenz et al. 2015).

Rather than competing with this well-established and useful software, `loadflex` works

seamlessly with `rloadest`. In particular, the `loadReg2` function in `loadflex` can be wrapped around the `loadReg` function from `rloadest` to form a load model that is built on `LOADEST` code but provides the same simple interface as other models in the `loadflex` suite (Fig. 1B). Adding the `loadflex` interface makes it possible to generate point predictions from `rloadest` models (e.g., Fig. 3C) separately from the aggregation of those point predictions to longer time periods, permitting faster analyses and bringing transparency to one black-box attribute of standard `rloadest` models. Further, the interface makes `rloadest` models compatible with the composite method, bringing an entirely new layer of functionality to `rloadest` models without sacrificing any of the careful thought and long experience that has gone into their design.

Composite model.—The composite method is most appropriate when regression predictions are available but show intermediate-term biases, i.e., predictions that are consistently too low or consistently too high over periods for which one or more observations are available and autocorrelated. The method can only be applied to time periods for which some observations of river chemistry exist, and the method thus cannot be used for extrapolation to unobserved periods of the past or future. Within those constraints, the composite method excels in making full use of the available data and in generating accurate predictions at the temporal resolution of the predictors (e.g., 15-minute discharge), which should lead to more accurate predictions at the monthly to annual scales.

A composite method model begins with a regression model, which in `loadflex` may take either of the forms described above (`loadReg2` or `loadLm`). In our demonstration we begin with `no3_reg2`, the `loadReg2` load model fitted in Fig. 1B and displayed in Fig. 3C. The regression model is then used to make predictions at a series of time points for which observations are available. The predictions and observations are compared to generate a set of "residuals", which differ from standard regression residuals in that they may be computed from data that differ in whole or part from the original calibration data. These residuals may be computed as absolute differences (observation − prediction) or relative differences (observation/prediction; absolute differences shown in Fig. 3E).

The key premise of the composite method is that an interpolation among these residuals accurately describes the temporal pattern of errors in the regression predictions. For example, Fig. 3E suggests that the regression model consistently underestimates concentrations in February and overestimates them in June and July of 2003, such that it would be appropriate to make an upward correction in February and a downward correction in June and July. To achieve such changes systematically, the composite method corrects the regression predictions (at the resolution of the predictor dataset, e.g., daily or hourly discharge observations) by the residuals as interpolated to that same resolution (Fig. 3F). This is done by adding the absolute residuals to the regression predictions or multiplying the predictions by the relative residuals (Fig. 3G). The resulting corrected predictions (Fig. 3D) can improve on pure interpolation predictions because their fine temporal structure reflects any variation in the regression predictions (e.g., a storm pulse in discharge and corresponding pulse or dip in concentrations). They can improve on pure regression predictions because their longer-term magnitudes are more accurate (e.g., corrected for any multi-week anomalies in the relationship between concentration and the regression predictors, such as a summer of unusually low concentrations).

The dataset used to fit the regression model must meet the assumptions of the regression method, including a lack of autocorrelation among residuals whenever uncertainty estimates are required. Consequently, the regression calibration dataset should usually have no finer than a monthly to weekly resolution (Lorenz et al. 2015). The interpolation phase, in contrast, not only permits but actually depends on the autocorrelation of residuals; otherwise, the correction would only add noise to the regression predictions (Aulenbach 2013). It is therefore possible in `loadflex` to perform the interpolation phase using a second dataset, which may have higher resolution than the regression calibration data if additional observations are available. Separating the regression and interpolation datasets has not been recommended in any earlier studies, to our knowledge, and has theoretical benefits in its ability to satisfy both

the independence of regression data and the autocorrelation of interpolation data. This separation can also make the prediction process more flexible by permitting the interpolation model to be fitted to a separate time period of data than that of the regression model. Lastly, the separation can make the prediction process more computationally efficient when the period spanned by the regression dataset is long but the period of interest for prediction is short; in this case, interpolation need only be done for that shorter period.

Interpolation of composite-method residuals is implemented in `loadflex` using an object created by the `loadInterp()` function (the same described earlier) but fitted not to raw concentrations or fluxes but to the residuals. Most interpolation methods available for `loadInterp` models are equally applicable to raw observations or to residuals—in particular, the piecewise linear method has been the standard since the early descriptions of the composite method (Huntington et al. 1994, Aulenbach and Hooper 2006). Additional methods were developed specifically for use with the composite method and have specific advantages in that application; one such method is the triangular interpolation proposed by Verma et al. (2012). The `loadflex` package also offers new interpolation options not previously employed in the composite method literature, including a distance-weighted function that causes the residual correction to approach zero as the time of the prediction point gets farther away from the time of the nearest observation point. In addition, `loadflex` allows the interpolation method to be applied to either absolute or relative residuals (an idea advanced by Verma et al. 2012), and to those residuals in log or linear space. Choices between proportional and absolute, between log and linear, and among the six implemented interpolation methods yield $2 \times 2 \times 6 = 24$ pre-defined options for interpolation of residuals in `loadflex`. In addition to these pre-defined interpolation functions, others may be defined by the user; the only requirement is that the user-defined function accepts arguments for the dates and values of the residuals to interpolate among and a new set of dates for which the interpolation should be made.

While all of the above options are made available to the user, we see reason to prefer the absolute, log-space option and have made this the default. In non-log space, relative interpolations have outcompeted absolute interpolations in tests to date (Verma et al. 2012). Absolute methods in log space produce quite similar predictions to those of relative interpolations, and they have the added advantage of operating in the same space in which the regression model was fitted. Based on our results and those of Verma et al. (2012), we use the absolute, log-space option as a default. However, the best interpolation function for a given dataset will depend on the researcher's objectives and the nature of that particular dataset. For simplicity of interpretation, a linear or rectangular interpolation could be best. For datasets where the researcher expects a smooth change in the concentration-discharge relationship during the time between two calibration observations, a relative interpolation could be important. For datasets with slow changes in the C-Q relationship (e.g., from one season to the next), a linear or smooth spline interpolation could be most effective, while for datasets with rapid changes in that relationship (e.g., between stormflow and baseflow), a distance-weighted interpolation might be best.

Point predictions

Point predictions, i.e., predictions of concentration or flux that each correspond to a row of predictor data, can be generated from any load model in a single call to the `predictSolute()` function (Fig. 1C). The inputs to this function include the fitted load model and a dataset of predictors. The output is a vector or table of predictions. For convenience, the predictions take whichever format is specified by the user (flux rate or concentration), even when the underlying regression model makes predictions in another format, and they are reported in linear space even if the regression model makes raw predictions in log space.

Transformation bias correction.—Linear regression model equations for load estimation nearly always have a log transformation on the left-hand side. Because researchers are generally interested in the non-logged estimates of concentration or flux rate, the model predictions must be retransformed by exponentiation; however, this introduces a bias, usually downward, in the

predictions (Ferguson 1986, Koch and Smillie 1986). Each exponentiated prediction should therefore be multiplied by a bias correction factor to reduce or eliminate this bias (Cohn et al. 1989). Several parametric solutions are available for regression models; for example, `rloadest` uses an adjusted maximum likelihood estimator (AMLE) by default (Cohn et al. 1989, Runkel et al. 2004, Cohn 2005).

Alternatives to bias correction are to use hyperbolic functions (Johnson et al. 1969), nonlinear models (Asselman 2000) or generalized linear models (Cox et al. 2008, Wang et al. 2011) rather than using log-linear models in the first place. However, the relative simplicity of log-linear models gives them ongoing value for the foreseeable future.

In the current version of `loadflex`, linear regression models use a simple error correction based on the theoretical relationship between lognormal and normal distributions: the final, retransformed prediction m is computed as

$$m = \exp\left(\mu + \frac{\sigma_p}{2}\right)$$

where μ is the log-space prediction from the model and σ_p is the standard error of prediction (SEP) for that point. Our composite method implementation currently uses a composite of error corrections as needed: the intermediate, regression-based predictions are bias-corrected according to the type of model employed (e.g., `rloadest` models use the AMLE estimator), and the residual-corrected predictions are either retransformed with the same method as for linear regressions if the residuals correction is done in log space, or left in linear space if the residuals correction is done in that space.

Uncertainty estimation.—There is uncertainty associated with each point prediction from any type of model, and `loadflex` estimates this uncertainty along with the prediction if requested. The format of this uncertainty is specified by the arguments `se.fit` and `se.pred` (each TRUE or FALSE) and `interval` ("none," "confidence," or "prediction") in the call to `predictSolute()`. Most of the work in estimating this uncertainty is actually completed during the earlier phase of model fitting, and the methods used are specific to the type of model being fit.

For interpolation models, uncertainty is estimated in `loadflex` by delete-one jackknife, a process closely related to N-fold cross validation (Hastie et al. 2009). In the jackknife process, the interpolation model is refit to a series of datasets that each differ from the original by the omission of exactly one observation. The difference between the omitted observation and the prediction at that point is then squared to estimate one error, and the average of the errors across all N of the refit models is interpreted as the mean squared prediction error for the model. The standard error of prediction (SEP) is the square root of this term (Kunsch 1989, Cressie 1993). If requested, prediction intervals are computed by the percentile method (Fox 2008), i.e., as the 2.5% and 97.5% quantiles of a normal distribution with the mean set to the predicted value and the standard deviation set to the SEP.

For regression models, the standard error of prediction is computed as the square root of the sum of coefficient uncertainty (the variance of the fit) plus the residual error variance (estimated as the mean of the squared residuals). Equations for extracting coefficient uncertainty and prediction error from standard regression models are well known (Freund and Wilson 2003). The comparable equations for censored data are slightly more complicated and are handled well within the `rloadest` package (Runkel et al. 2004, Cohn 2005).

In contrast to methods for regression models, no methods for uncertainty estimation for the composite method have been historically available (Aulenbach 2013). Within `loadflex`, we have implemented an algorithm for estimating composite method uncertainty that is itself a composite of two approaches: we employ a parametric bootstrap (Efron 1979, Cressie 1993) where the parameters are the covariances of the regression model coefficients and the statistic of interest is the error after the interpolation phase as estimated by delete-one jackknife (Kunsch 1989, Cressie 1993). First, the coefficients of the regression model are resampled from their multivariate normal distribution to get a new set of plausible coefficients. This parametric resampling process is comparable to, but more computationally efficient than, the non-parametric alternative of refitting the model to resampled calibration data. (The non-parametric option is also available for composite models containing

`loadLm` regression models.) New predictions are then made at the interpolation calibration points, from which the residuals are calculated by comparison to the original observations. Those new residuals are then employed in a jackknife process where during each iteration, all but one residual is used to create an interpolation, the interpolated corrections are reapplied to the regression predictions (from the resampled coefficients) to determine a final prediction, and the error is computed as the difference between the original observation and this final prediction. Each element of the current interpolation dataset is left out in turn to produce a set of errors. The entire bootstrap process is repeated many times, each time going through the parametric coefficient resampling, the inner jackknife loop of residuals correction and prediction, and calculation of the mean squared error for that inner loop. The mean squared errors from all bootstrap iterations are then averaged to find the estimated variance for the full composite model. The square root of that variance is the SEP. The number of bootstrap repetitions can be adjusted by the user to achieve any desired balance between computation time and convergence. For the particular application summarized in Fig. 1, we found that the default of 100 repetitions gave an error estimate that was within ~1% of the error estimated with 1000 repetitions.

For practical and theoretical reasons, the space (log or linear) in which composite method errors are computed depends on the space in which the interpolation was done. Interpolations done in linear space have the potential to yield negative predictions of flux or concentration at some time points; consequently, the most sensible space for computing the mean squared error (MSE) is also linear. This is consistent with the assumption implied by the user in choosing to interpolate residuals in linear space—if the user expects normally distributed errors about the regression predictions, then an interpolation in linear space will make the most sense. In contrast, an interpolation in log space both (1) produces entirely positive predictions, making it practical to compute an MSE in log space, and (2) implies that the user expects roughly lognormal distributions of errors about the regression predictions. For these reasons, interpolations in log space are paired with overall uncertainty estimates in log space, and linear with linear. All predictions generated by `predictSolute()` are then retransformed as needed so that they can be reported in linear space.

The approach we have implemented to estimate composite method uncertainty, combining a parametric bootstrap with a jackknife, should provide a reasonable upper bound on the uncertainty for the composite model for two reasons. First, the parametric resampling of regression coefficients allows us to consider uncertainty in those coefficient estimates without introducing random noise, which would reduce the autocorrelation of the regression residuals. The entire motivation behind the interpolation phase is to make use of autocorrelation in those residuals, so it is imperative that the autocorrelative structure of the residuals be preserved. Second, a delete-one jackknife approach is appropriate for a dataset of evenly spaced observations to be interpolated. In each iteration, the deleted observation is maximally distant in time from the observations being used for interpolation. In other words, any other point to be predicted by the composite method model in the prediction phase will be closer to an observation than the observations being left out in the jackknife process, and is therefore likely to have less error than that predicted by the jackknife. Consequently, the delete-one process outlined above is unlikely to underestimate error in the composite method. At the same time, we expect that the overestimation will not be dramatic. We delete just one observation at a time so that in any given iteration, the point being left out is being predicted by the regression model and by all of the neighboring points. A delete-*k* approach would sometimes have more than one adjacent point being left out, introducing additional and irrelevant error because those points would be corrected for more distant residuals than is the case in the actual composite method prediction algorithm.

Diagnostics of model fit

Any model that is fit should also be evaluated. Traditional metrics and methods of model evaluation apply to either type of regression model currently available through `loadflex` (`loadLm` or `loadReg2`). For example, users will usually want to inspect the coefficient estimates

and p values for model terms and metrics of fit such as the adjusted R^2. These can be accessed through summaries of the fitted models contained within the `loadLm` or `loadReg2` objects. Those inner fitted models can be extracted using the `getFittedModel()` function, which returns an object that can be summarized using the standard `print()` and `summary()` functions (Fig. 1D). Models fitted by the USGS `loadReg()` function (i.e., those contained within `loadReg2` objects) also supply variance inflation factors, bias metrics, and a suite of diagnostic plots (Lorenz et al. 2015). To inspect the regression model contained within a composite model, the user can apply `getFittedModel()` to the `loadLm` or `loadReg2` object before passing it to `loadComp()`.

Users will also want to inspect the residuals from any fitted model to look for patterns in the mean or variance of those residuals. Residuals may be extracted from regression models by standard methods appropriate to `lm` or `loadReg` classes after obtaining the inner model with `getFittedModel()`. Alternatively, `loadflex` supplies the `getResiduals()` method, which can be applied to any type of load model defined by `loadflex` and which returns residuals in many possible formats, including flux or concentration, absolute (predicted − observed) or relative (predicted/observed), and log or linear space. Obtaining log-space or relative residuals is particularly useful when considering whether or not to apply the composite method to a regression model, because it is often desirable to apply the residuals interpolation and correction using logged or relative residuals. The residuals obtained by `getResiduals()` or other methods may then be plotted versus time, day of year, observed value, discharge, etc. to look for heteroskedasticity and any patterns uncaptured by the current model formulation (Fig. 1D).

A regression model requires the independence (non-autocorrelation) of residuals to accurately assess uncertainty, while an interpolation or composite model requires that the autocorrelation of concentrations or fluxes is strong enough to reasonably extrapolate forward and backward from an observation (Aulenbach 2013). `loadflex` supplies two convenience functions, `residDurbinWatson()` and `estimateRho()`, to test whether the autocorrelations of the calibration and interpolation residuals meet those assumptions. For example, in the case of the `loadReg2` model named `no3_reg2` in Fig. 1, the Durbin-Watson d statistic is 1.32 and the autocorrelation ρ is 0.34 for the calibration residuals, whereas $d = 1.00$ and $\rho = 0.50$ for the interpolation residuals. These statistics confirm that we have achieved the desired difference in autocorrelation between the two datasets.

An especially useful metric for composite models is the fraction of total flux accounted for by the residuals correction as opposed to the underlying regression model predictions. Based on several related metrics employed by Aulenbach and Hooper (2006), here we define a slightly modified metric, the Correction Fraction (CF), as follows:

$$\mathrm{CF} = \frac{\sum_i |R_i| \Delta t}{\sum_i L_i \Delta t}$$

where $\sum$ indicates a sum over all predictions, L_i is the ith composite prediction of flux or concentration, R_i is the ith interpolated residual in the same format (flux or concentration, non-log space) as L_i, and Δ_i is the time period represented by the ith prediction. This metric can be computed in `loadflex` by a call to the `getCorrectionFraction()` function (Fig. 1D). A high value of CF indicates that the composite model is relying heavily on the interpolated residuals rather than on the regression predictions and that the user should consider whether a better regression model is possible for use within the composite method. For example, the composite model which employs the `no3_reg2` regression model (`no3_comp`; Fig. 1B), has a concentration correction fraction of 0.21, whereas a composite model employing the weaker `no3_lm` regression model would have had a flux correction fraction of 0.32, indicating a heavier reliance on the interpolation phase to compensate for inaccuracies in the regression predictions.

Load aggregation

Loads can be aggregated from point predictions to values of mean concentration, mean flux rate, or total flux at daily, monthly, annual, or other intervals using the `aggregateSolute()` function (Fig. 1E). Fundamentally, aggregation always involves a sum of the constituent point estimates; in the case of mean concentration or

mean flux rate, the sum is followed by division by the number of observations. At present `loadflex` requires that the user has filled in any data gaps before passing point predictions to the aggregation function; future work may add more sophisticated gap-handling methods to the package.

The uncertainty in an aggregated estimate builds on the uncertainty in point estimates but also involves the propagation of error through the summing process (Cohn 2005, Lehrter and Cebrian 2010). Classical propagation of error for a total load ($\sum L$) dictates that the variance in that total load is equal to the sum of the variance-covariance matrix of the point estimates contributing to the total, as follows:

$$\mathrm{var}(\sum_i L_i) = \sum_{i,j} \Sigma[i,j] = \sum_{i,j} \mathrm{cov}(L_i, L_j)$$

where large $\sum$ indicates a sum and small Σ is the variance-covariance matrix relating variation in each point load estimate L_i to variation in each other point estimate L_j, i.e., describing the correlation of errors among the L_i. The contents of Σ usually cannot be determined from a calibration dataset because the resolution of that dataset is often too coarse to identify the autocorrelation of errors at scales of hours to days. Instead, `loadflex` allows the user to specify an assumption based on an expert understanding of the river and the chosen model. Given a specification of the structure of the correlation matrix, `loadflex` automatically calculates the covariance for each pair of errors (i.e., the value of each cell in Σ) as the product of their correlation and variances.

Several options for specifying Σ are predefined in `loadflex`. One option is consistent with that assumed by the `LOADEST` and `rloadest` approaches: the correlation in errors between each pair of points can be set at 1 when those points fall on the same calendar date and 0 when their dates differ. Another option is a smooth band of perfect correlation (i.e., 1) within a time interval of the user's choosing, with perfect independence (i.e., 0) outside that time interval. Still another represents a first-order autocorrelation process such that the correlation between two points declines smoothly as a function of the points' distance in time at a rate determined by the coefficient ρ. These assumptions are all oversimplifications, but they vary in realism, computation time, and similarity to the assumptions used in other load models (e.g., `LOADEST`), so the user may prefer one above the others for a given application.

A first-order autocorrelation coefficient, ρ, is required to fully specify one of the above options and may be valuable in evaluating any of the options. The value of ρ must often be chosen based on the user's expert judgment. However, the helper function `estimateRho()` can also be used to produce an empirical estimate of ρ if passed a supplemental dataset of high-resolution concentration and discharge observations. This supplemental dataset should be one for which the user is confident that prediction errors will share the same autocorrelation structure as those from the original estimation dataset. For example, one or two years of high-resolution sensor-based data might have a ρ value that is representative of the same site for several preceding years; in the case of the Lamprey River sensor data, `estimateRho()` finds $\rho = 0.9996$ per 15 minutes or $\rho = 0.9634$ per day for residuals of the `no3_reg2` regression model. Estimating ρ precisely requires substantially more high-resolution data than are available at most sites, and `loadflex` therefore also allows the user to arbitrarily specify the value of ρ for the purpose of defining an error autocorrelation matrix.

A separate question in the aggregation process is which distribution to assume for flux or concentration estimates and their errors. Point estimates are usually assumed to follow lognormal distributions, as noted earlier. For aggregate concentration and flux estimates, however, the mean and standard error are computed in linear space, consistent with the central limit theorem that the sum of a large number of random variables, no matter their individual distributions, is itself normally distributed. This assumption is increasingly questionable for increasingly small time periods, an issue that neither we nor our predecessors have addressed. In most cases, however, the aggregate loads of interest—monthly or annual, for example—will have ample observations to justify the application of the central limit theorem.

Like the mean estimates, the confidence intervals around aggregate estimates may be

computed in `loadflex` using an assumption of normality; however, the default algorithm follows `LOADEST` and Cohn (2005) in computing those intervals based on a lognormal distribution. Specifically, the confidence or prediction interval around a mean estimate μ with standard error σ in linear space has a corresponding mean μ_L and standard error σ_L in log space:

$$\sigma_L^2 = \log_e\left(1 + \frac{\sigma^2}{\mu^2}\right)$$

$$\mu_L^2 = \log_e(\mu) - \frac{\sigma_L^2}{2} = \log_e\left(\frac{\mu^2}{\sqrt{\mu^2 + \sigma^2}}\right)$$

which indicate the following 95% confidence interval bounds I:

$$I_L = \mu_L \pm 1.96\sigma_L$$
$$I = \exp(\mu_L \pm 1.96\sigma_L).$$

Using a lognormal distribution in this way has the advantage of producing confidence bounds that stay above zero and are thus more realistic than those produced for a normal distribution.

Biogeochemical application

The `loadflex` interface makes it easy to fit multiple models to the same data. The logical next step is to compare the fitted models and predictions. We demonstrate this process in the context of 12 years of weekly nitrate measurements and quarter-hourly observations of discharge in the Lamprey River, New Hampshire. A motivating observation for analyzing these data was that a large storm in October 2005 and two exceptionally large floods in May 2006 and April 2007 seem to have altered nitrate concentrations and fluxes in subsequent years. A lasting shift in nitrate concentrations would indicate a fundamental change in watershed nitrate delivery or retention processes, whereas a shift in nitrate fluxes could be important to the nitrate-sensitive eelgrass beds of the Great Bay Estuary downstream (Trowbridge 2012). We can use monthly fluxes and concentrations to quantitatively evaluate (Q1) the short-term effects of each flood on nitrate loads for the month and (Q2) the longer-term effects of the three events on nitrate loads in subsequent years.

To illustrate the full range of models that may be constructed with `loadflex`, we fitted one model from each of four model types (a `loadInterp` model named `no3_interp`, a `loadLm` named `no3_lm`, a `loadReg2` named `no3_reg2`, and a `loadComp` named `no3_comp`), using fitting arguments and formulas as in Fig. 1. We created greater contrast between the two regression model examples by using a simple regression model formula for `no3_lm` ($\ln C = a_0 + a_1 \ln Q$) and a more complex formula for `no3_reg2` (`model(9)`: $\ln(CQ) = a_0 + a_1 \ln Q + a_2 \ln Q^2 + a_3 \sin(2\pi t) + a_4 \cos(2\pi t) + a_5 t + a_6 t^2$), where a_0–a_6 are model coefficients, C is concentration, Q is discharge, and t is decimal time. However, our biogeochemical questions require us to address the possibility of stepwise shifts in concentration or flux, and neither of these regression equations contains terms for such shifts. To give each model a maximal chance of detecting such shifts while retaining the contrast among model structures, we fitted each model three times to three successive periods: the five water years preceding the hydrologic perturbations (10/1/2000–9/30/05; Before), the two water years of large floods and storms (10/1/05–9/30/07; During), and the five following water years (10/1/07–9/30/12; After). We then made predictions for each period, combined them into a single time series per model, and aggregated the predictions to monthly averages. Fitting each model to three periods permitted a fairer comparison between the regression models, which are constrained in their representation of permanent shifts in concentration or flux, and the interpolation-based methods, which freely adapt to any changes over time.

Our first question was whether the two months of major flooding differed in their mean concentrations or total fluxes. To evaluate the hypothesis that these months were not significantly different, we applied z-tests (equivalent to t-tests with normally distributed statistics) to the estimates and uncertainties of flux and concentration for those two months. We applied a separate z-test to the output of each load model to explore how the choice of load model influences the answer to this biogeochemical question.

Our second question was whether monthly mean concentrations or fluxes differed systematically among the water years that preceded,

contained, or followed the three large hydrologic events. To answer this question we fitted a new simple linear model to the monthly concentration or flux values. We will call this the "flood model" to distinguish it from the four load models that produced the monthly estimates. In this flood model, we accommodated the annual patterns in concentration and flux by fitting a categorical term for Month; the fitted values of these Month coefficients are unimportant to our question and are not reported here. The other term in the model equation was a categorical variable indicating whether the monthly estimate falls in the Before, During, or After period (water years 2001–05, 2006–07, or 2008–12, respectively). Coefficients fit to this term indicate whether there are significant differences between periods. We fitted a total of eight flood models to look for differences in either concentrations or fluxes as estimated by each of the four load models.

Model performance assessments

When high-resolution data are available, several quantitative metrics are available to aid the user in evaluating and comparing models. We demonstrate the use of such metrics by analyzing the sensor-based data from the Lamprey River, NH, containing two years of 15-minute resolution observations of discharge and nitrate. While recognizing that in situ sensors, like laboratory instruments, have measurement error, we used these sensor-based estimates of nitrate concentration as the best available approximation of the "true" concentrations at each 15-minute interval for the purpose of evaluating the four models demonstrated in this manuscript (Fig. 1).

To assess the models using these new data, we first subsampled the full sensor dataset to simulate a more typical sampling regime of approximately weekly grab samples. We set the first "grab sample" in this simulated dataset to occur on a random day and time in the 1-week interval beginning on 9/24/12, selected by random number generator from a uniform distribution across that period. We then fixed that day and time of week as the target sampling time for every week, but also simulated typical field challenges by selecting the actual sampling time for each week as a time normally distributed around the target sampling time with a standard deviation of 0.6 days. The result was a series of 104 weekly samples across the sampling period, with some realistic variation in the precise day and time of each sample. This weekly dataset was used to fit the interpolation model and the interpolation step of the composite model. To reduce the presence of autocorrelation in the datasets used to fit the regression models (including the regression model component of the composite model), we further subsampled the simulated dataset to 70% of its original values by leaving out every third or fourth weekly observation. Subsampling reduced autocorrelation in the residuals from 0.57 to 0.43 for the `no3_lm` model and from 0.35 to 0.14 for the `no3_reg2` model. Autocorrelated residuals have no direct effect on model fit but cause regression models to underestimate their uncertainty; consequently, the standard error for the log(discharge) term increased from 0.051 to 0.061 for the `no3_lm` model and from 0.060 to 0.074 for the `no3_reg2` model with subsampling.

From the simulated weekly dataset and its 70% subsample, we next fitted the four models as in Fig. 1B and compared the predictions at 15-minute resolution and the aggregated predictions at monthly resolution to the same "true" values as observed in the complete, 15-minute resolution dataset of concentration and discharge measurements. The "true" monthly concentration values were computed by aggregation from the 15-minuted data using the same function, `aggregateSolute()`, that we used to generate aggregate values from model predictions. We produced "true" fluxes at the 15-minute resolution by computing the product of concentration, discharge, and a units conversion factor; these were then aggregated as usual to get the monthly "true" fluxes. Using those true values, for each fitted model we assessed the accuracy, bias, and precision of predictions and the accuracy of the uncertainty interval by computing the following metrics.

The relative root mean squared error (RRMSE) describes the accuracy of the predictions. RRMSE was computed from the relative differences between predictions and observed values as

$$\mathrm{RRMSE} = \sqrt{\frac{1}{N}\sum_{i}^{N}\left(\frac{V_{P,i} - V_{O,i}}{V_{O,i}}\right)^2}$$

where N is the total number of 15-minutely or monthly observations, $V_{P,i}$ is the ith element in

the vector of predicted values (either flux or concentration), and $V_{O,i}$ is the ith element in the vector of observed values.

The bias (B) of the predictions describes any consistent directional deviation of predictions from observations. B was computed as the median difference between each predicted and observed value.

$$B = \text{median}(V_{P,i} - V_{O,i})$$

The average relative interval length (ARIL) summarizes the uncertainty reported by a model by normalizing the uncertainty interval sizes by the observed value at each point, then taking the average of the normalized sizes (Jin et al. 2010, Vigiak and Bende-Michl 2013). We computed ARIL as the average of the relative 95% prediction interval lengths

$$\text{ARIL} = \frac{1}{N}\sum_{1}^{N}\frac{L_{P,high,i} - L_{P,low,i}}{V_{O,i}}$$

where $L_{P,high,i}$ is the upper 95% prediction interval bound for the ith element and $L_{P,low,i}$ is the lower bound.

Lastly, the bracketing frequency (BF) measures the accuracy of the reported uncertainty intervals. We computed BF as the frequency with which the observed value falls within the prediction interval at each time point in the set of predictions:

$$\text{BF} = \frac{1}{N}\sum_{i}^{N}\begin{cases} 1, L_{P,low,i} \le V_{O,i} \le L_{P,high,i} \\ 0, \text{otherwise} \end{cases}.$$

All four metrics were calculated for each fitted model, for predictions of both flux and concentration, and at both the 15-minute and the monthly resolution of prediction. We then repeated the entire process for a total of 100 iterations. In each iteration, we resampled the 2-year sensor record to simulate a new weekly dataset, fitted the models to that dataset (with further subsampling to obtain the calibration data for the regression models, as above), and computed the metrics. We report the means of the metrics produced from those 100 iterations to ensure that our metrics are not unduly influenced by a quirk of any one specific subset of the available data.

Results

Biogeochemical observations

Whether predicted at 15-minute intervals (Fig. 4) or aggregated to monthly time steps (Fig. 5), nitrate fluxes show stronger annual periodicity than do nitrate concentrations. The difference can be attributed to the dominance of discharge over concentration in determining patterns in flux: The observed discharge has an annual period (Fig. 2), and the `no3_lm` model predicts nearly constant concentrations over time and yet predicts roughly periodic fluxes (Fig. 4, row 2). The greatest fluxes reliably occur in the winter and spring, with inter-annual variability in the precise timing and magnitude of those fluxes. In contrast, concentration patterns are less readily characterized and appear to follow roughly one to two oscillations per year.

As can be seen in the uncertainty intervals in Figs. 4 and 5, flux predictions tend to be more precise than concentration predictions. This is largely because discharge is used as a multiplier to obtain flux and can be precisely observed. Monthly predictions have narrower confidence intervals than point predictions due to the increase in confidence that is gained by summing over a large number of predictions whose autocorrelation is low, or at least assumed to be so.

We asked whether concentrations in the river and/or fluxes toward the Great Bay estuary differed significantly between the two months of major flooding (circled in Fig. 6). Because the four load models (`no3_interp`, `no3_lm`, `no3_reg2`, and `no3_comp`) each make different predictions for monthly fluxes and their uncertainties, the results of the z-tests also differ (Table 1). Results based on predictions from the regression models, `no3_lm` and `no3_reg2`, indicate no significant difference in concentration between the two months, and either a non-significant or weakly significant difference in flux. In contrast, the `no3_interp` and `no3_comp` models indicate significantly higher concentrations and fluxes in the second flood (greater by 0.037–0.040 mg/L and 77–81 kg/d). The models differ most dramatically in their estimates of absolute monthly concentrations and fluxes: For example, in May '06 `no3_comp` estimates a mean concentration of 0.09 mg/L while `no3_lm` estimates 0.17 mg/L, and

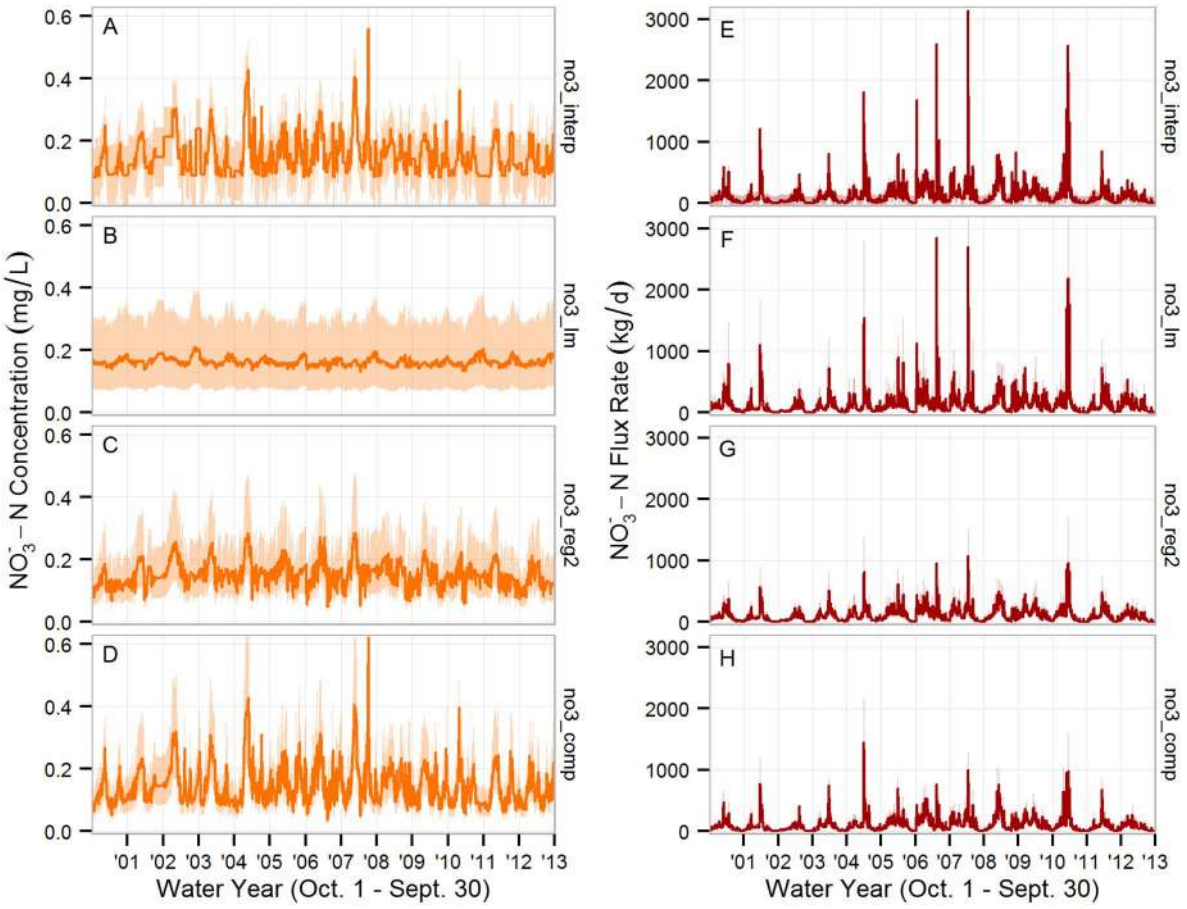

Fig. 4. Point predictions and 95% prediction intervals from each of the four models (labels on right) for concentration (A–D) and flux rate (E–H).

in April '07 `no3_comp` estimates a mean flux rate of 244 kg/d while `no3_interp` estimates 514 kg/d (Table 1, Fig. 6).

We also looked for differences among the multi-year periods preceding, containing, and following the three large floods and storms. As with the two-month comparison above, the results depend on the load model used to generate the monthly mean concentrations or fluxes (Table 2). The four load models all support our initial qualitative observation that concentrations were lower after the floods than before (decline of 0.013–0.021 mg/L, or 10–15%). However, of the four load models, only the regression models `no3_lm` and `no3_reg2` indicate that concentrations in the years containing the floods and storms increased significantly from corresponding months in the preceding years (increase of 0.008–0.012 mg/L, or 4–9%). The models that include interpolation, `no3_interp` and `no3_comp`, indicate changes of similar size but no significance. With respect to flux rates, the four models agree that fluxes were substantially higher in the storm and flood years than in the preceding years, but the magnitude of this difference ranges almost two-fold from 54 kg/d (`no3_comp`) to 102 kg/d (`no3_lm`). The models also disagree on whether fluxes shifted after the storm and flood years, with only the `no3_reg2` model indicating a significant change (an increase of 25 kg/d) relative to the years before the floods.

Model performance assessments

Performance metrics based on the nitrate sensor data indicate substantial differences

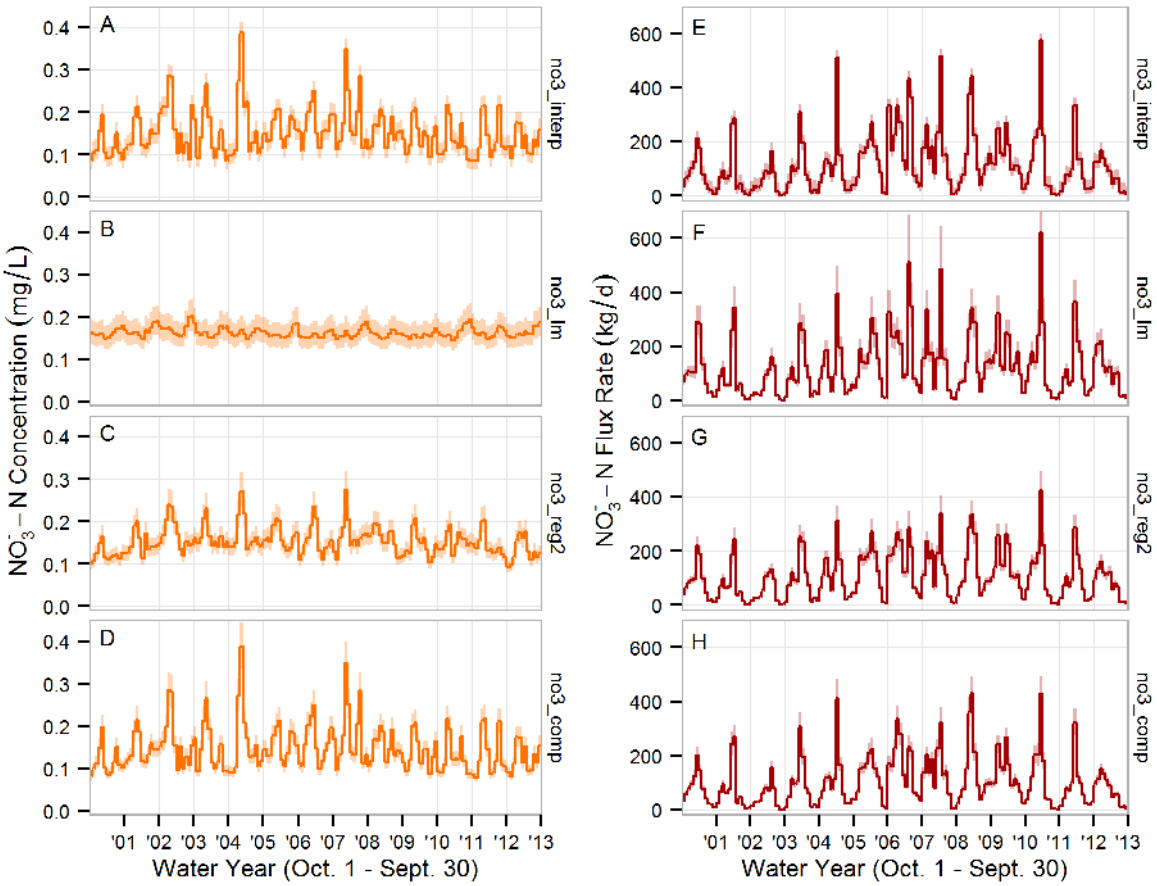

Fig. 5. Aggregate monthly predictions with 95% uncertainty intervals from each of the four models (labels on right) for concentration (A–D) and flux rate (E–H).

among the four models in both accuracy and bias (Table 3). The accuracy of predictions at 15-minute intervals is substantially better for the `no3_interp` and `no3_comp` models (RRMSE = 0.3 for each) than for the `no3_lm` and `no3_reg2` models (RRMSE = 1.1 and 0.7, respectively). All four models are more accurate for monthly estimates (RRMSE of 0.20–0.54) than for 15-minute estimates (0.3–1.1 as above), illustrating the utility of aggregation in smoothing over short-term noise to find the longer-term signal.

The bias is a median error and is therefore units-specific, such that concentration biases can readily be compared to one another but should be expected to be consistently smaller than flux biases because the range of observed concentrations (0.08–0.557 mg/L) is much smaller than the range of observed fluxes (1.8–1440 kg/d). Within each category, the models including interpolation (`no3_interp` and `no3_comp`) are less biased than the regression-only models (`no3_lm` and `no3_reg2`), sometimes by as much as two orders of magnitude (Table 3).

The average relative interval length (ARIL) summarizes the size of the uncertainty intervals reported by each load model. Reported precision at 15-minute resolution is better for the `no3_comp` model (ARIL = 1.9) than for any other model (ARIL = 2.6–5.3). The `no3_comp` and `no3_reg2` models report more precise monthly predictions than the `no3_lm` and `no3_interp` models (0.59–0.61, compared to 0.67–1.30).

Bracketing frequency (BF) measures the accuracy of each model's self-reported uncertainty intervals. A bracketing frequency of 0.95 would indicate perfect agreement between the reported uncertainty (as a 95% PI) and the error rate observed for a particular prediction task. The

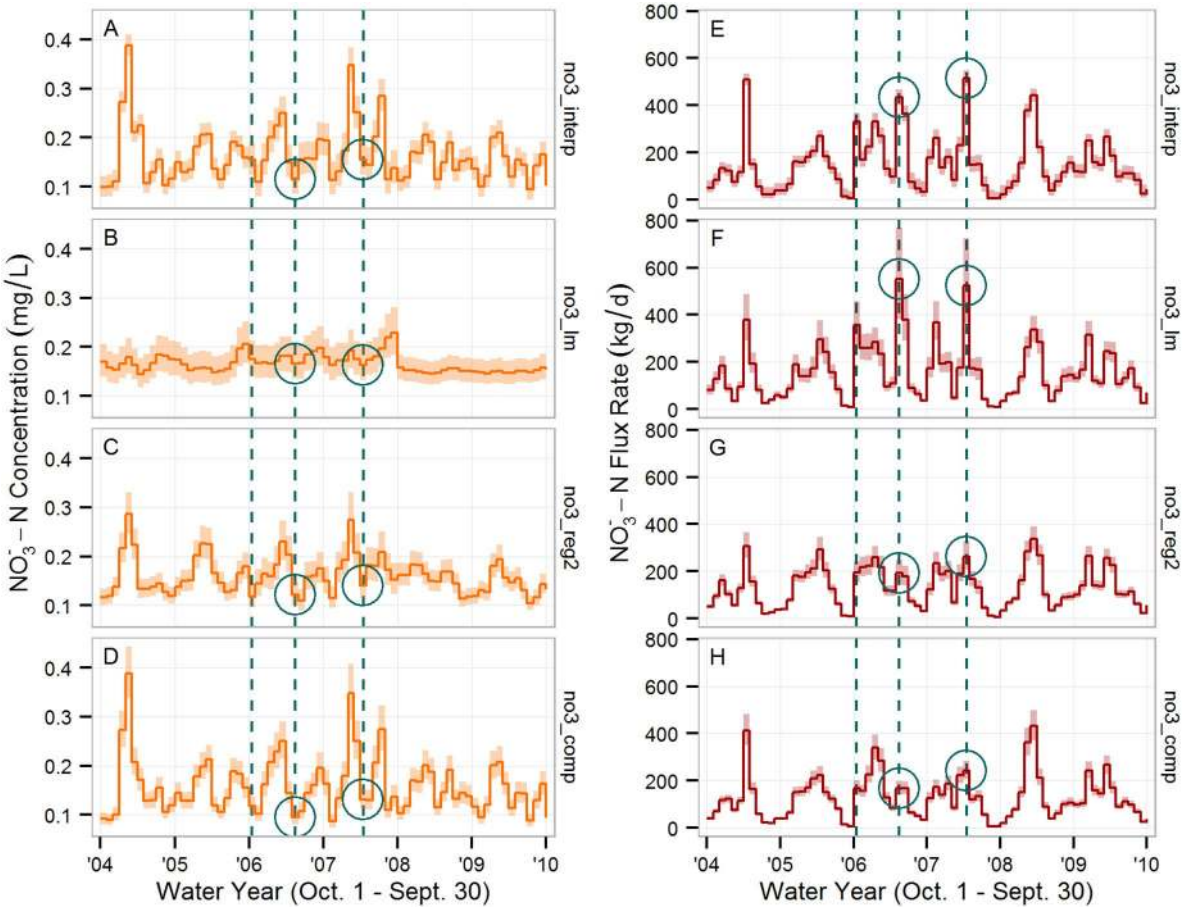

Fig. 6. Aggregate monthly predictions for the seven years before, during, and after the large storm of October 2005 (water year 2006) and two large floods of May 2006 and April 2007. The storm and flood months are indicated with blue vertical lines. The two flood months for which we directly compared fluxes by z-test are circled in blue. The Before (2001–2006), During (2006–2008), and After (2008–2013) periods for the fixed effects model are delineated by the x-axis tick marks at the water years 2006 and 2008.

predictions at 15-minute resolution tend to fall close to this target, with the regression models `no3_lm` and `no3_reg2` coming closest (BF = 0.951 and 0.957) and the interpolation and composite approaches reporting uncertainties somewhat larger than necessary (BF = 0.991 and 0.986, respectively). That the interpolation and composite methods report conservative

Table 1. Monthly estimates and results of z-tests assessing the difference in monthly mean concentration or flux between the 100-year floods of May 2006 and April 2007.

	NO_3^--N concentration (mg/L)				NO_3^--N flux rate (kg/d)			
Model	May 2006	April 2007	z stat	p value	May 2006	April 2007	z stat	p value
no3_interp	0.11	0.15	1.67	<0.05	433	514	3.31	<0.001
no3_lm	0.17	0.16	−0.10	n.s.	551	524	−0.19	n.s.
no3_reg2	0.12	0.14	0.98	n.s.	192	263	1.71	<0.05
no3_comp	0.09	0.13	2.72	<0.01	167	244	3.16	<0.001

Note: The z statistic and p values are from z-tests employing the mean and standard error of each flood month as estimated by each of the four load models, with a positive z statistic indicating a higher value in April 2007 than in May 2006.

Table 2. Results of simple linear regressions on load model outputs, testing for differences in monthly concentration or flow before, during, and after the major storms and floods of October 2005, May 2006, and April 2007.

	NO_3^--N concentration (mg/L)				NO_3^--N flux rate (kg/d)			
Model	Before	During-Before	After-Before	R^2	Before	During-Before	After-Before	R^2
no3_interp	0.14	0.017 (n.s.)	−0.021 (<0.01)	0.51	56	94 (<0.0001)	27 (n.s.)	0.41
no3_lm	0.18	0.008 (<0.01)	−0.021 (<0.0001)	0.65	66	102 (<0.0001)	31 (n.s.)	0.40
no3_reg2	0.13	0.012 (<0.05)	−0.013 (<0.01)	0.53	45	57 (<0.0001)	25 (<0.05)	0.55
no3_comp	0.13	0.013 (n.s.)	−0.020 (<0.01)	0.58	40	54 (<0.001)	23 (n.s.)	0.48

Notes: Model equation was Value ~ Month + Phase, where Month is a categorical variable for October through the following September, and Phase is one of Before October 2005, During October 2005–September 2007, or After September 2007. Statistical models were fit to load model outputs for months from October 2000 to September 2012. The intercepts in the columns titled "Before" are for September means in the years Before October 2005. The values in the "During-Before" columns are the amount by which concentration or flux in the During period exceeds that in the Before period, followed by the p-values of those estimates in parentheses. Similarly, the "After-Before" columns indicate the amount by which the After period exceeds the Before period. Adjusted R^2 values of the statistical models are given for each load model and prediction format.

(large) uncertainty intervals is, as reasoned in the Methods section, expected. When applied to interpolations, the delete-1 jackknife produces less conservative uncertainty than any delete-*k* alternative; nonetheless, it evaluates each model based on performance at test points that are more distant in time than any points in the actual set of predictions. This stringent test yields a 95% uncertainty interval that may be larger than required. It is thus notable that the `no3_comp` ARIL is often the smallest of the four models: Even by this potentially conservative measure of uncertainty, the composite method approach has reduced the uncertainty in estimates of flux and concentration relative to the competing models.

At the monthly resolution the interpolation and composite methods still have bracketing frequencies close to the 0.95 target (BF = 0.944–0.978). In contrast, the regression-only models have substantially lower bracketing frequencies (0.582–0.676); this difference reflects the relatively high bias and autocorrelation of errors in the regression model predictions, which leads to an underestimate of uncertainty for monthly fluxes. The interpolations in the `no3_interp` and `no3_comp` models reduce the bias and autocorrelation of errors such that the uncertainty estimates are accurate even after aggregation.

Discussion

The large number of approaches used to estimate watershed fluxes and concentrations can be explained by the fact that each method has a different set of strengths and weaknesses (Letcher et al. 1999, Johnes 2007). For this reason, the `loadflex` package makes four major ap-

Table 3. Metrics of performance for each of the four demonstrated load models.

		RRMSE		Bias		ARIL		BF	
Prediction format	Model	15 min	Month	15 min	Month	15 min	Month	15 min	Month
Concentration	no3_interp	0.3	0.34	−0.00002	0.0003	2.7	0.67	0.991	0.968
	no3_lm	1.1	0.47	0.01068	0.0272	2.8	0.77	0.951	0.582
	no3_reg2	0.7	0.31	0.01362	0.0103	2.3	0.59	0.957	0.575
	no3_comp	0.3	0.20	0.00034	−0.0004	1.9	0.60	0.986	0.973
Flux rate	no3_interp	0.3	0.35	−0.01	0.2	5.3	1.30	0.986	0.978
	no3_lm	1.1	0.54	4.04	4.1	2.8	0.85	0.951	0.594
	no3_reg2	0.7	0.28	4.83	1.7	2.3	0.61	0.957	0.676
	no3_comp	0.3	0.20	0.13	−1.8	1.9	0.61	0.986	0.944

Notes: Models are as fit in Fig. 1. Metrics are for predictions of either concentration or flux at 15-minute or monthly resolution. RRMSE: relative root mean squared error. Bias: median difference between predicted and observed value. ARIL: average of the relative 95% prediction interval lengths. BF: bracketing frequency, or the frequency with which the observed value falls within the prediction interval at that point. Each cell contains the average over 100 iterations of resampling the two-year dataset, refitting the models, and computing the metrics.

proaches equally accessible by the same interface to facilitate model fitting and comparison. Interpolation models are simple to explain to other researchers and make no requirements about the discharge-concentration relationship, but they may fail to capture important concentration changes during storms, and they often require more data than regression models (Robertson and Roerish 1999). Linear regression models are more transparent than `LOADEST` or `rloadest` models, making them a good choice for initial exploration of the data and the modeling options. `LOADEST`/`rloadest` models are well documented and frequently applied, highly appropriate when the data include censored chemistry observations, and pre-programmed with several commonly successful model formulas. However, both simple linear models and those implemented by `LOADEST`/`rloadest` are susceptible to violations of key assumptions that can lead to biased predictions (Hirsch 2014). The composite method is less well established, more complex than interpolation or regression models, dependent on data availability during the time period of interest, and resistant to traditional methods of uncertainty estimation; however, predictions by the composite method can be less biased and more precise than predictions by other methods.

In the application of these four models to the Lamprey River, New Hampshire, USA, we found several a priori reasons to favor the composite method over the alternatives. The weekly chemistry data were sufficient in resolution, quality, and time span for use with any of the modeling approaches. The 7-parameter `loadReg2` model, `no3_reg2`, explained a non-trivial proportion of variance in the concentrations ($R^2 = 0.38$), arguing in favor of a regression-based approach rather than a simpler interpolation-only model. However, there were also observable intermediate-term biases in the regression predictions, such as the underestimates in February and overestimates in June and July of the 2003 water year (Fig. 3C). Lastly, there was a detectable autocorrelation of the residuals when all observations were used (autocorrelation: $\rho = 0.48$; Durbin-Watson statistic: $d = 1.03$), from which we hypothesized that a composite of regression predictions and interpolated residuals could yield good predictions at multiple time scales.

Model tests using two years of high-resolution chemistry data supported our initial preference for the composite method as a way to generate accurate and precise predictions (Table 3). The `no3_interp` and `no3_comp` models yielded smaller RRMSEs and much smaller biases than the regression alternatives. Another study in review for this special issue also found comparable performance of interpolation (period-weighted) and composite approaches for NO_3^- fluxes in several watersheds; in contrast, composite models typically outperformed interpolation and regression models for SO_4^{2-}, Si, and DOC (Aulenbach et al. unpublished manuscript). The inability of composite models to improve on interpolation for NO_3^- suggests that regression on discharge and/or season provides little useful information about sub-weekly variation in NO_3^- concentrations. Composite models may nonetheless have an advantage over other models with respect to uncertainty: `no3_comp` had the lowest reported uncertainty (ARIL) of all models for 15-minute predictions and tied with `no3_reg2` for the lowest ARIL for monthly predictions. Unlike the regression models, composite method uncertainty estimates are also reliable at multiple resolutions, with bracketing frequencies (BF) remaining close to the target for both 15-minute and monthly estimates.

Model evaluation using sensor-based chemistry data (as in Table 3) will not be possible for all rivers or time periods, owing to the recent development and nontrivial cost of nutrient sensors. When possible, however, such evaluations can inform the final choice of a modeling approach and the interpretation of concentration or flux estimates. Models should be compared by multiple objective metrics, such as those demonstrated here, to assess the models' performance both in predicting loads and in estimating the uncertainty around those predictions.

The choice of modeling approach has implications for model predictions and subsequent analyses of those predictions. As we showed through two simple applications, a researcher's ability to test for differences or trends in solute loads depends on the quality and precision of the concentration or flux estimates. In a z-test comparing the months of peak flux in the years just before and just after a year of large storms, a significant difference in concentration was only

detected with predictions from two of the four load models (Table 1). Similarly, a fixed-effects flood model relating monthly fluxes to their occurrence before, during, or after two years of hydrologic disturbances also gave different results for each set of load model predictions (Table 2). These contrasting results illustrate the sensitivity of load analyses to the choice of a load estimation model, and they emphasize the importance of making an informed choice in matching each new study site and dataset to the most appropriate load model. Although the burden of that choice must ultimately fall on the researcher, `loadflex` can ease the process by supplying a uniform interface for fitting and assessing several of the most common load models. Future work may incorporate still more load modeling approaches into the `loadflex` package.

A key objective of our project was to lower the barriers that have prevented past researchers from choosing the best available model and quantifying uncertainties in the chosen method. The `loadflex` package is free, open-source, and extensible, and it joins the ranks of many other ecology-relevant packages now available in the R statistical language (Kneib and Petzoldt 2007). Our package provides smooth integration with the `LOADEST`/`rloadest` estimation framework and implements options for interpolation, linear regression, and composite models, all of which are accessible via a simple and uniform interface. The implementation of uncertainty estimation for all four model types makes it possible to compare models and report results with greater rigor. Our hope is that `loadflex` will make more methods readily available to more researchers, facilitating model comparison and application to achieve the best possible load estimates across a wide range of sampling regimes and watersheds.

Acknowledgments

The authors wish to thank Kristofor Voss for statistical guidance, Jody Potter for data collection and management, Michelle Daley for Lamprey sample collection and land cover analyses, and Adam Wymore, Jeff Taylor, and two anonymous reviewers for thoughtful critiques of the manuscript. Funding for the project was provided by the National Science Foundation (EPS-1101245 and EAR-1331841). Additional support was provided by the USDA National Institute of Food and Agriculture McIntire-Stennis Project 1006760 and the NH Agricultural Experiment Station; this is Scientific Contribution Number 2632. Funding for collection of the example data set from the Lamprey River was provided by the EPA (USEPA Cooperative Agreement R-83058601-0), NH Water Resources Research Center, NH Agricultural Experiment Station, NH Sea Grant, USGS, and NSF (IIA-1330641).

Literature Cited

Asselman, N. E. M. 2000. Fitting and interpretation of sediment rating curves. Journal of Hydrology 234:228–248.

Aulenbach, B. T. 2013. Improving regression-model-based streamwater constituent load estimates derived from serially correlated data. Journal of Hydrology 503:55–66.

Aulenbach, B. T., H. T. Buxton, W. A. Battaglin, and R. H. Coupe. 2007. Streamflow and nutrient fluxes of the Mississippi-Atchafalaya river basin and subbasins for the period of record through 2005: US Geological Survey Open-File Report 2007-1080. http://toxics.usgs.gov/pubs/of-2007-1080/index.html

Aulenbach, B. T., and R. P. Hooper. 2006. The composite method: an improved method for stream-water solute load estimation. Hydrological Processes 20:3029–3047.

Birgand, F., C. Faucheux, G. Gruau, B. Augeard, F. Moatar, and P. Bordenave. 2010. Uncertainties in assessing annual nitrate loads and concentration indicators: Part 1. Impact of sampling frequency and load estimation algorithms. Transactions of the ASABE 53:437–446.

Bowes, M. J., and W. A. House. 2001. Phosphorus and dissolved silicon dynamics in the River Swale catchment, UK: a mass-balance approach. Hydrological Processes 15:261–280.

Buso, D. C., G. E. Likens, and J. S. Eaton. 2000. Chemistry of precipitation, streamwater, and lakewater from the Hubbard Brook Ecosystem Study: a record of sampling protocols and analytical procedures. General Technical Report NE-275. USDA Forest Service, Northeastern Research Station, Newtown Square, Pennsylvania, USA.

Cohn, T. A. 2005. Estimating contaminant loads in rivers: an application of adjusted maximum likelihood to type 1 censored data. Water Resources Research 41:W07003.

Cohn, T. A., D. L. Caulder, E. J. Gilroy, L. D. Zynjuk, and R. M. Summers. 1992. The validity of a simple statistical model for estimating fluvial constituent loads: an empirical study involving nutrient loads entering Chesapeake Bay. Water Resources Research 28:2353–2363.

Cohn, T. A., L. L. Delong, E. J. Gilroy, R. M. Hirsch, and D. K. Wells. 1989. Estimating constituent loads. Water Resources Research 25:937–942.

Cox, N. J., J. Warburton, A. Armstrong, and V. J. Holliday. 2008. Fitting concentration and load rating curves with generalized linear models. Earth Surface Processes and Landforms 33:25–39.

Cressie, N. A. C. 1993. Statistics for spatial data. Revised edition. Wiley, New York, New York, USA.

Efron, B. 1979. Bootstrap methods: another look at the jackknife. Annals of Statistics 7:1–26.

FEMA [Federal Emergency Management Agency]. 2008. Independent evaluation of recent flooding in New Hampshire. Federal Emergency Management Agency, Washington, D.C., USA.

Ferguson, R. I. 1986. River loads underestimated by rating curves. Water Resources Research 22:74–76.

Fox, J. 2008. Applied regression analysis and generalized linear models. Second edition. SAGE, Thousand Oaks, California, USA.

Freund, R. J., and W. J. Wilson. 2003. Statistical methods. Second edition. Academic Press, Elsevier Science, Burlington, Massachusetts, USA.

Grimm, N. B. 1987. Nitrogen dynamics during succession in a desert stream. Ecology 68:1157–1170.

Groffman, P. M., N. L. Law, K. T. Belt, L. E. Band, and G. T. Fisher. 2004. Nitrogen fluxes and retention in urban watershed ecosystems. Ecosystems 7:393–403.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. The elements of statistical learning: data mining, inference, and prediction. Second edition. Springer, New York, New York, USA.

Hirsch, R. M. 2014. Large biases in regression-based constituent flux estimates: causes and diagnostic tools. Journal of the American Water Resources Association 50:1401–1424.

Hirsch, R. M., and L. A. De Cicco. 2015. User guide to exploration and graphics for RivEr Trends (EGRET) and dataRetrieval: R packages for hydrologic data. Version 2.0. U.S. Geological Survey, Reston, Virginia, USA.

Hirsch, R. M., D. L. Moyer, and S. A. Archfield. 2010. Weighted regressions on time, discharge, and season (WRTDS), with an application to Chesapeake Bay river inputs. Journal of the American Water Resources Association 46:857–880.

Hruška, J., P. Krám, W. H. McDowell, and F. Oulehle. 2009. Increased dissolved organic carbon (DOC) in Central European streams is driven by reductions in ionic strength rather than climate change or decreasing acidity. Environmental Science and Technology 43:4320–4326.

Huntington, T. G., R. P. Hooper, and B. T. Aulenbach. 1994. Hydrologic processes controlling sulfate mobility in a small forested watershed. Water Resources Research 30:283–295.

Jin, X., C.-Y. Xu, Q. Zhang, and V. P. Singh. 2010. Parameter and modeling uncertainty simulated by GLUE and a formal Bayesian method for a conceptual hydrological model. Journal of Hydrology 383:147–155.

Johnes, P. J. 2007. Uncertainties in annual riverine phosphorus load estimation: impact of load estimation methodology, sampling frequency, baseflow index and catchment population density. Journal of Hydrology 332:241–258.

Johnson, N. M., G. E. Likens, F. H. Bormann, D. W. Fisher, and R. S. Pierce. 1969. A working model for the variation in stream water chemistry at the Hubbard Brook Experimental Forest, New Hampshire. Water Resources Research 5:1353–1363.

Kneib, T., and T. Petzoldt. 2007. Introduction to the special volume on "Ecology and ecological modeling in R." Journal of Statistical Software 22:1–7.

Koch, R. W., and G. M. Smillie. 1986. Bias in hydrologic prediction using log-transformed regression models. Journal of the American Water Resources Association 22:717–723.

Kunsch, H. R. 1989. The jackknife and the bootstrap for general stationary observations. Annals of Statistics 17:1217–1241.

Lehrter, J. C., and J. Cebrian. 2010. Uncertainty propagation in an ecosystem nutrient budget. Ecological Applications 20:508–524.

Letcher, R. A., A. J. Jakeman, W. S. Merritt, L. J. McKee, B. D. Eyre, and B. Baginska. 1999. Review of techniques to estimate catchment exports. Technical Report 99/73. New South Wales Environment Protection Authority, Sydney, New South Wales, Australia.

Likens, G. E., F. H. Bormann, N. M. Johnson, D. W. Fisher, and R. S. Pierce. 1970. Effects of forest cutting and herbicide treatment on nutrient budgets in the Hubbard Brook watershed-ecosystem. Ecological Monographs 40:23–47.

Likens, G. E., F. H. Bormann, R. S. Pierce, J. S. Eaton, and N. M. Johnson. 1977. Biogeochemistry of a forested ecosystem. Springer, New York, New York, USA.

Lorenz, D., R. Runkel, and L. De Cicco. 2015. rloadest: river load estimation. U.S. Geological Survey, Mounds View, Minnesota, USA.

McCrackin, M. L., J. A. Harrison, and J. E. Compton. 2014. Factors influencing export of dissolved inorganic nitrogen by major rivers: a new, seasonal, spatially explicit, global model. Global Biogeochemical Cycles 28:269–285.

Miller, C. R. 1951. Analysis of flow-duration, sediment-rating curve method of computing sediment yield. U.S. Bureau of Reclamation, Denver, Colorado, USA.

Moatar, F., and M. Meybeck. 2005. Compared perfor-

mances of different algorithms for estimating annual nutrient loads discharged by the eutrophic River Loire. Hydrological Processes 19:429–444.

Moyer, D. L., R. M. Hirsch, and K. E. Hyer. 2012. Comparison of two regression-based approaches for determining nutrient and sediment fluxes and trends in the Chesapeake Bay watershed. Scientific Investigations Report 2012-5244. U.S. Geological Survey, Richmond, Virginia, USA.

New Hampshire Department of Environmental Services. 2013. Lamprey River water management plan. Watershed Management Bureau, New Hampshire Department of Environmental Services, Concord, New Hampshire, USA.

NH GRANIT [New Hampshire Geographically Referenced Analysis and Information Transfer System]. 2011. Impervious surfaces in coastal NH and southern York County ME-2010. Complex Systems Research Center, University of New Hampshire, Durham, New Hampshire, USA.

NOAA [National Oceanic and Atmospheric Administration]. 2006. Coastal change analysis program regional land cover. Office for Coastal Management, Washington, D.C., USA.

Peters, N. E., J. B. Shanley, B. T. Aulenbach, R. M. Webb, D. H. Campbell, R. Hunt, M. C. Larsen, R. F. Stallard, J. Troester, and J. F. Walker. 2006. Water and solute mass balance of five small, relatively undisturbed watersheds in the U.S. Science of the Total Environment 358:221–242.

Porterfield, G. 1972. Computation of fluvial-sediment discharge. Techniques of water-resources investigations of the United States Geological Survey. U.S. Geological Survey, Arlington, Virginia, USA.

Preston, S. D., V. J. Bierman, and S. E. Silliman. 1989. An evaluation of methods for the estimation of tributary mass loads. Water Resources Research 25:1379–1389.

Prokushkin, A. S., O. S. Pokrovsky, L. S. Shirokova, M. A. Korets, J. Viers, S. G. Prokushkin, R. M. W. Amon, G. Guggenberger, and W. H. McDowell. 2011. Sources and the flux pattern of dissolved carbon in rivers of the Yenisey basin draining the Central Siberian Plateau. Environmental Research Letters 6:045212.

Raymond, P. A., and J. E. Saiers. 2010. Event controlled DOC export from forested watersheds. Biogeochemistry 100:197–209.

Robertson, D. M., and E. D. Roerish. 1999. Influence of various water quality sampling strategies on load estimates for small streams. Water Resources Research 35:3747–3759.

Runkel, R. L. 2013. Revisions to LOADEST, April 2013. U.S. Geological Survey, Reston, Virginia, USA.

Runkel, R. L., C. G. Crawford, and T. A. Cohn. 2004. Load estimator (LOADEST): a FORTRAN program for estimating constituent loads in streams and rivers. USGS Techniques and Methods Book 4, Chapter A5. U.S. Geological Survey, Reston, Virginia, USA.

Stenback, G. A., W. G. Crumpton, K. E. Schilling, and M. J. Helmers. 2011. Rating curve estimation of nutrient loads in Iowa rivers. Journal of Hydrology 396:158–169.

Thomas, R. B. 1988. Monitoring baseline suspended sediment in forested basins: the effects of sampling on suspended sediment rating curves. Hydrological Sciences Journal 33:499–514.

Toor, G. S., R. D. Harmel, B. E. Haggard, and G. Schmidt. 2008. Evaluation of regression methodology with low-frequency water quality sampling to estimate constituent loads for ephemeral watersheds in Texas. Journal of Environment Quality 37:1847.

Trowbridge, P. 2012. Assessments of aquatic life use support in the Great Bay Estuary for chlorophyll-a, dissolved oxygen, water clarity, eelgrass habitat, and nitrogen. Technical support document. New Hampshire Department of Environmental Services, Concord, New Hampshire, USA.

Trowbridge, P., M. A. Wood, J. T. Underhill, and D. S. Healy. 2014. Great Bay nitrogen non-point source study. State of New Hampshire, Department of Environmental Services, Concord, New Hampshire, USA.

U.S. Census Bureau. 2011. 2010 Census of population and housing. http://www.census.gov/prod/www/decennial.html

Vanni, M. J., W. H. Renwick, J. L. Headworth, J. D. Auch, and M. H. Schaus. 2001. Dissolved and particulate nutrient flux from three adjacent agricultural watersheds: a five-year study. Biogeochemistry 54:85–114.

Verma, S., M. Markus, and R. A. Cooke. 2012. Development of error correction techniques for nitrate-N load estimation methods. Journal of Hydrology 432–433:12–25.

Vigiak, O., and U. Bende-Michl. 2013. Estimating bootstrap and Bayesian prediction intervals for constituent load rating curves. Water Resources Research 49:8565–8578.

Walker, W. W. 1996. Simplified procedures for eutrophication assessment and prediction: user manual. U.S. Army Engineer Waterways Experiment Station, Vicksburg, Mississippi, USA.

Wang, Y.-G., P. Kuhnert, and B. Henderson. 2011. Load estimation with uncertainties from opportunistic sampling data: a semiparametric approach. Journal of Hydrology 396:148–157.

Webb, B. W., J. M. Phillips, D. E. Walling, I. G. Littlewood, C. D. Watts, and G. J. L. Leeks. 1997. Load estimation methodologies for British rivers and their relevance to the LOIS RACS(R) programme. Science of the Total Environment

194–195:379–389.
Worrall, F., N. J. K. Howden, and T. P. Burt. 2013. Assessment of sample frequency bias and precision in fluvial flux calculations: an improved low bias estimation method. Journal of Hydrology 503:101–110.
Yanai, R. D., et al. 2015. Sources of uncertainty in estimating stream solute export from headwater catchments at three sites. Hydrological Processes 29:1793–1805.

Supplemental Material

Ecological Archives

The Supplement is available online: http://dx.doi.org/10.1890/ES14-00517.1.sm