Received: 20 July 2020 | Revised: 28 November 2020 | Accepted: 30 November 2020
DOI: 10.1002/hyp.14000

RESEARCH ARTICLE

WILEY

Predicting high-frequency variation in stream solute concentrations with water quality sensors and machine learning

Mark B. Green[1,2] | Linda H. Pardo[2] | Scott W. Bailey[2] | John L. Campbell[2] | William H. McDowell[3] | Emily S. Bernhardt[4] | Emma J. Rosi[5]

[1]Department of Earth, Environmental, and Planetary Sciences, Case Western Reserve University, Cleveland, Ohio

[2]Northern Research Station, U.S. Forest Service, North Woodstock, New Hampshire

[3]Department of Natural Resources and the Environment, University of New Hampshire, Durham, New Hampshire

[4]Department of Biology, Duke University, Durham, North Carolina

[5]Cary Institute of Ecosystem Studies, Millbrook, New York

Correspondence
Mark B. Green, Department of Earth, Environmental, and Planetary Sciences, Case Western Reserve University, 10900 Euclid Ave, Cleveland, OH 44106-7078.
Email: mbg78@case.edu

Funding information
Directorate for Biological Sciences, Grant/Award Numbers: 1256696, 1637685; Office of Integrative Activities, Grant/Award Number: 1101245

Abstract

Stream solute monitoring has produced many insights into ecosystem and Earth system functions. Although new sensors have provided novel information about the fine-scale temporal variation of some stream water solutes, we lack adequate sensor technology to gain the same insights for many other solutes. We used two machine learning algorithms – Support Vector Machine and Random Forest – to predict concentrations at 15-min resolution for 10 solutes, of which eight lack specific sensors. The algorithms were trained with data from intensive stream sensing and manual stream sampling (weekly) for four full years in a hydrologic reference stream within the Hubbard Brook Experimental Forest in New Hampshire, USA. The Random Forest algorithm was slightly better at predicting solute concentrations than the Support Vector Machine algorithm (Nash-Sutcliffe efficiencies ranged from 0.35 to 0.78 for Random Forest compared to 0.29 to 0.79 for Support Vector Machine). Solute predictions were most sensitive to the removal of fluorescent dissolved organic matter, pH and specific conductance as independent variables for both algorithms, and least sensitive to dissolved oxygen and turbidity. The predicted concentrations of calcium and monomeric aluminium were used to estimate catchment solute yield, which changed most dramatically for aluminium because it concentrates with stream discharge. These results show great promise for using a combined approach of stream sensing and intensive stream discrete sampling to build information about the high-frequency variation of solutes for which an appropriate sensor or proxy is not available.

KEYWORDS
biogeochemistry, machine learning, stream solutes, water quality

1 | INTRODUCTION

Stream chemistry provides insights into a wide range of ecosystem and Earth system dynamics. Because catchment processes with high spatial variability are integrated into a catchment outlet sample, considerable advances in understanding fundamental processes such as mineral weathering, element limitation of vegetation growth, and soil development can be made using stream chemistry and the whole catchment approach (Bormann & Likens, 1979). Catchment solute yields vary in response to changing atmospheric deposition, shifting climate, ecosystem succession, extreme weather events, and major forest disturbance (e.g., Bernhardt et al., 2003; Eshleman &

 © 2020 John Wiley & Sons Ltd

Sabo, 2016; Valett et al., 2002). The resulting patterns in the timing and magnitude of solute fluxes have important implications for water supply management and stream ecosystem dynamics (Allan & Ibañez Castillo, 2009). Continuous, long-term monitoring of atmospheric inputs and stream solute outputs using the small catchment ecosystem approach has allowed us to understand changes in ecosystem element storage (e.g., Bailey, Buso, & Likens, 2003), the influence of nitrogen deposition on catchment nutrient retention (Bettez et al., 2015), and exploration of biogeochemical stationarity (Basu et al., 2010).

Advances in water quality sensing are providing data on solute concentration in streams and rivers at higher temporal resolution than the typical weekly sampling that has been the backbone of many small catchment studies (Likens, 2013; Schleppi et al., 2006; Sebestyen et al., 2011). Stream water solute concentrations vary over diurnal (Aubert & Breuer, 2016; Neal et al., 2013; Pellerin et al., 2009) and storm scales (Inserillo et al., 2017; Koenig et al., 2017), which are not well represented in monitoring programmes that rely on weekly or coarser water sampling. How estimates of annual element yields improve with higher frequency data remains to be seen. Pellerin et al. (2009) used nitrate and dissolved organic matter sensors to quantify the yield of both sensed solutes during a snowmelt season. They estimated that the fluxes calculated from high resolution data were 14% higher for dissolved organic carbon and 2% higher for nitrate compared to estimates derived from their interpolations of weekly sampling. For many solutes, high temporal resolution sensors have not yet been developed or have poor detection, making it difficult to gain the benefits of high-frequency sampling.

Stream solutes that cannot yet be directly sensed, can still be predicted by combining proxy data with predictive algorithms. Concentrations of mercury are known to vary with UV absorbance, and thus optical sensing has allowed better estimation of the temporal variation of stream water mercury concentrations (Dittman et al., 2009). High-frequency conductivity measurements were effective predictors of all major ions derived from weathering of mountaintop removal mined watersheds (Ross et al., 2018). High-frequency sulphate time series were produced with discharge as an input variable for multiple machine learning algorithms (Mewes et al., 2020). Kisi and Parmar (2016) predicted monthly chemical oxygen demand in an Indian river with nutrient and other water quality information. High-frequency stream water phosphorus has been predicted using turbidity as a proxy (Lannergård et al., 2019). Overall, these studies have accurately estimated water quality constituents.

Building on these previous studies, we used high-frequency water quality sensors in combination with an extensive set of discrete samples and two machine learning algorithms to predict non-sensed solute concentrations. We explored model predictions for all routinely monitored solutes, but highlight Ca^{2+} and Al^{3+} here because past research has demonstrated their sensitivity to anthropogenic acidification (Lawrence et al., 2015; Likens et al., 1996), and their response in surface waters as soils recover from acidification remains an important forest management issue. These two solutes also have distinctly different transport dynamics. Dissolved Ca^{2+} is generally derived from mineral weathering within the catchment, with transport modified by soil exchange and biotic cycling, resulting in a neutral-to-negative concentration relationship with discharge (Godsey et al., 2009; Johnson et al., 1969). In comparison, Al^{3+} is also derived from mineral weathering but its mobility is limited at moderate pH and by co-precipitation with organic matter in soil development. As a result, it is predominantly derived from eluvial, shallow-to-bedrock soils near the catchment divide that are only hydrologically connected to the stream network during high flow (Bailey et al., 2019). This results in a positive Al^{3+} concentration relationship with discharge (Lawrence et al., 1988). Differences in solute transport dynamics have resulted in differential importance of high-frequency information in estimating stream loads (Aulenbach et al., 2016; Swistock et al., 1997). We examined this difference by using the predicted Ca^{2+} and Al^{3+} concentrations to estimate their solute yields. This study presents an advancement in predicting high frequency stream solute concentrations (15-min). The application of two machine learning models with a rich set of discrete stream water samples covering 10 solutes and multiple years of high-frequency sensor data allowed us to demonstrate how high-frequency solute data sets might be produced and used in catchment solute budgets.

2 | METHODS

2.1 | Site

This study was conducted at the Hubbard Brook Experimental Forest, located in the White Mountains of New Hampshire, USA (latitude 43°56′ N, longitude 71°45′ W). Watershed 3 (W3) at Hubbard Brook is a reference catchment of mature (approximately 110 years old) mixed hardwood forest. Dominant tree species are sugar maple (*Acer saccharum*), American beech (*Fagus grandifolia*), and yellow birch (*Betula alleghaniensis*; Siccama et al., 2007). The catchment is a steep (mean slope = 17.1%), glaciated hillside of spodic soils that show spatial variation consistent with podzolization due to lateral groundwater flux (Bailey et al., 2014; Gillin et al., 2015). Hubbard Brook has a warm-summer humid continental climate according to the Köppen climate classification (Bailey, Hornbeck, et al., 2003; Kottek et al., 2006).

2.2 | Sensor data

Specific conductance (SC), pH, fluorescent dissolved organic matter (FDOM), turbidity, dissolved oxygen (DO), and NO_3^- were sensed at the outlet of W3 between October 2012 and January 2017. A Yellow Springs Instruments EXO2 multi-parameter sonde generated the SC, pH, FDOM, turbidity, and DO data, and a Satlantic submersible UV analyser generated the NO_3^- data. All sensor data were logged at 15-min intervals. Snyder et al. (2018) describes the details of the deployment, maintenance, and quality assurance.

2.3 | Stream chemistry data

In total, 478 samples were collected during the study period (October 2012 to May 2017), including routine weekly samples and multiple storm and seasonal snow melt sampling campaigns. Immediately after collection, samples were passed through a pre-combusted (450°C) glass-fibre filter (0.7 μm nominal pore size). Samples were stored frozen prior to analysis, except for cation analyses (Al^{3+}, Ca^{2+}, Mg^{2+}, Na^{+}, K^{+} and Si), in which case an aliquot was poured off and refrigerated. Samples were shipped to the United States Department of Agriculture Forest Service, Forest Sciences Laboratory in Durham, New Hampshire where they were analysed for SO_4^{2-}, NO_3^{-} and Cl^{-} using ion chromatography (Metrohm 761); Ca^{2+}, Mg^{2+}, Na^{+}, K^{+} and Si with inductively coupled plasma optical emission spectroscopy (Agilent 730); dissolved organic carbon (DOC) with combustion catalytic oxidation on a total organic carbon analyser (Shimadzu TOC-V); and total monomeric Al (Al^{3+}) with the pyrocatechol violet method on a Flow Injection Analysis System (Lachat Quickchem). Precision and detection limits are given by USDA Forest Service, Northern Research Station (2019).

2.4 | Data analysis

Support vector machine regression (SVM) was used to predict the concentration of lab-measured solutes using sensed water quality data. Support Vector Machine learning algorithms are designed to identify patterns for use in prediction, and have been applied in water quality analyses (Kisi & Parmar, 2016; Tan et al., 2012). The SVM regression algorithm is described in depth by Vapnik (2013) and Drucker et al. (1997). Briefly, SVM regression involves fitting a hyperplane to the multi-dimensional data set, with a margin surrounding the hyperplane within which the cost is zero to the objective function. Correlated independent variables do not violate any assumptions in the SVM method. We used a radial kernel function to account for non-linearity in the data set. We set the Cost parameter to 1 and the gamma parameter for the radial kernel to 0.143 (the inverse of the number of variables or dimensions included in the model), which are default values in the algorithm. If the SVM regression predicted negative concentrations (this was the case only for NO_3^{-}), the prediction was set to zero.

We also used the random forest (RF) algorithm to predict the lab-measured solutes. The Random Forest model builds an ensemble (forest) of regression trees by making many random subsets of the data and using random variable sets to build many predictions of the dependent variable. Those many predictions are averaged to produce the final estimated dependent variable. Like SVM, the RF allows inclusion of correlated variables and can represent non-linear relationships. The RF model has been applied to problems ranging from water source identification to time series gap filling in biogeochemistry (e.g., Baudron et al., 2013; Kim et al., 2020). Further details about the RF model are described in Breiman (2001). We accepted default parameters for the RF model, including the number of trees required for the ensemble (n = 500) and the number of variables tried at each split in an individual tree (mtry = 2). We chose the SVM and RF models because both have been previously applied in hydrological contexts with strong results (e.g., Kim et al., 2020; Mewes et al., 2020). The main difference between the two is the RF uses discrete predictions, which can help identify non-linear patterns, and the SVM is a continuous function.

We used the e1071 R package to build the SVM models (Meyer et al., 2019) and the randomForest R package to build the RF models (Liaw & Wiener, 2002). We trained the SVM and RF models to predict solutes that were not sensed using a random 66% sample of the discrete water samples paired with sensor data (FDOM, pH, SC, DO, turbidity) and sine-transformed day of the year (DOY) to help explain any residual variance due to seasonality $\left(\sin\left(\frac{DOY}{365.25\pi}\right)\right)$. This sine transformation made DOY values on either side of January 1, which are normally numerically distant (DOY of 1 vs. 365), of similar magnitude and thus more representative of seasonality. The models were then used to predict solute concentrations for the remaining 33% of samples. This training and prediction approach was designed to improve interpolation between discrete samples, as opposed to make predictions of solute concentrations into the future. Comparisons of predicted and measured solute concentrations were quantified with the Nash-Sutcliffe (NS) efficiency (Nash & Sutcliffe, 1970), which compares the data variance to the 1:1 line (as opposed to an ordinary least squares line, as is the case with the commonly used R^2). Values of NS can range from 1 (perfect predictions) to $-\infty$ (poor predictions) with values lower than 0 indicating that the mean is a better prediction than the modelled values. Because the NS efficiency was based on a random subset of training data, we calculated the NS efficiency 1000 times to quantify the distribution of possible values. We described the central tendency of the distribution of NS values with the median, and the 95% confidence as the range between the 2.5th and 97.5th percentiles. When we predicted the concentration for the full data set for time series visualization or flux calculations (described below), we used 10 random models to ensure that each discrete water sample was predicted multiple times, and the mean of the multiple predictions were used to compare with measured concentrations.

We further tested the accuracy of the SVM and RF concentration predictions using sensed NO_3^{-}. We trained the NO_3^{-} model using discrete samples (n = 382) the same way we performed the training for other solutes, and tested the prediction on the sensed data, which accurately represents lab-measured values (sensed = 1.07[measured] + 0.008; r^2 = 0.98; Snyder et al., 2018).

The sensitivity of the SVM and RF models to input variables was tested by building a model using all available data (no training or testing data split), removing an individual independent variable from the model, and then re-calculating new NS efficiency values. The magnitude of the NS decrease was used to identify the most sensitive independent variables. We also used partial dependence plots to visualize the relationships, within the multivariate machine learning models, between our focus solutes (Ca^{2+} and Al^{3+}) and the independent variables. We used the pdp package in R to construct the partial dependence plots (Greenwell, 2017).

The annual flux of Ca^{2+} and Al^{3+} was calculated for 2016 by multiplying the predicted concentration by stream discharge for each 15-min period. Both concentration and discharge were assumed to be constant for the 15-min period after measurement, changing in a stepwise fashion. These fluxes from the 15-min predicted concentration were compared to the monthly fluxes based on weekly water chemistry samples using the period-weighted approach described in Aulenbach et al. (2016). Only 2016 was used for comparison because it was the year with the most complete sensor record.

3 | RESULTS

Concentrations of major ions in our study stream were generally low (median conductivity 11 μS cm^{-1}), with Ca^{2+} and Na^{+} the cations found in highest concentrations (median = 0.68 and 0.92 mg/L respectively) and SO_4^{2-} the dominant anion (median = 0.94 mg S/L; Figure 1). DOC concentrations were generally low (median = 2.3 mg/L), with the highest observed value reaching 10 mg/L in 12 August 2016 on a rising limb of a hydrograph after a 48 mm rain event. Concentration distributions of K^{+}, Al^{3+}, and DOC showed the most positive skew, indicating occasional high concentrations, and Si and SO_4^{2-} showed the least amount of skew, indicating more stable concentrations.

The median Nash-Sutcliffe (NS) efficiencies of predicted versus observed concentrations ranged between 0.29 for Cl^{-} and 0.79 for Na^{+} from the SVM, and 0.35 for Cl^{-} and 0.78 for Na^{+} from the RF (Table 1). The NS values from the RF models were generally higher than from the SVM models. For example, median RF-based NS efficiencies were equal to or greater than the SVM-based values for 7 of the 10 solutes. The 95% confidence intervals were generally greater for solutes that had lower median NS efficiency. For example, Cl^{-} and K^{+} had the lowest median NS from both algorithms, and had a 95% confidence interval >0.3. Both Mg^{2+} and Na^{+} had the highest median NS values and 95% confidence intervals <0.2 for both algorithms. Both Al^{3+} and DOC were exceptions to this general trend, both having relatively high median NS values (NS > 0.6) and 95% confidence intervals >0.3.

Both algorithms underpredicted the highest observed Ca^{2+} and Al^{3+}concentrations (Figure 2). This bias was indicated by the slope of the ordinary least squares (OLS) fit line, which was greater than the 1:1 line (Figure 2). The SVM predictions showed more scatter around the OLS line.

Time series comparisons between our continuous prediction of solute concentration and measured values from our discrete samples highlight time periods when model predictions are poor (Figure 3). For example, the relatively high peak event concentrations of Ca^{2+} and Al^{3+} tend to be underpredicted by the RF model. The SVM predictions

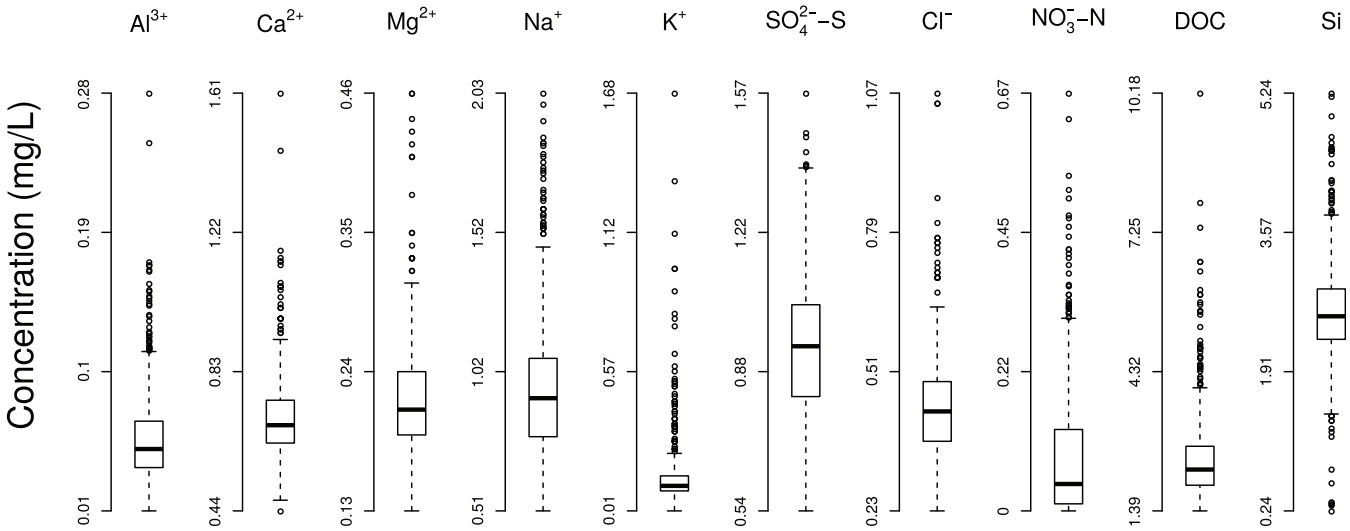

FIGURE 1 Boxplots showing the distributions of concentrations for each solute from this study. The box shows the median and the interquartile range and the whiskers extend to 1.5 times the interquartile range away from the box

TABLE 1 Nash-Sutcliffe efficiencies (NS) of the predicted versus measured solute concentration; 2.5th, 50th, and 97.5th percentile NS values from 1000 random test sets are shown for both the support vector machine and random forest models

Solute	Support vector machine NS (2.5th, 50th, and 97.5th percentiles)	Random forest NS (2.5th, 50th, and 97.5th percentiles)
Al^{3+}	0.41, 0.63, 0.86	0.48, 0.70, 0.85
Ca^{2+}	0.53, 0.64, 0.73	0.57, 0.66, 0.75
Mg^{2+}	0.68, 0.76, 0.82	0.71, 0.78, 0.84
Na^{+}	0.68, 0.79, 0.85	0.66, 0.77, 0.84
K^{+}	0.33, 0.49, 0.68	0.28, 0.56, 0.68
SO_4^{2-}	0.58, 0.74, 0.84	0.59, 0.74, 0.84
Cl^{-}	0.15, 0.29, 0.47	0.15, 0.35, 0.48
NO_3^{-}	0.72, 0.83, 0.89	0.68, 0.77, 0.85
DOC	0.34, 0.61, 0.78	0.33, 0.60, 0.79
Si	0.50, 0.65, 0.78	0.58, 0.70, 0.80

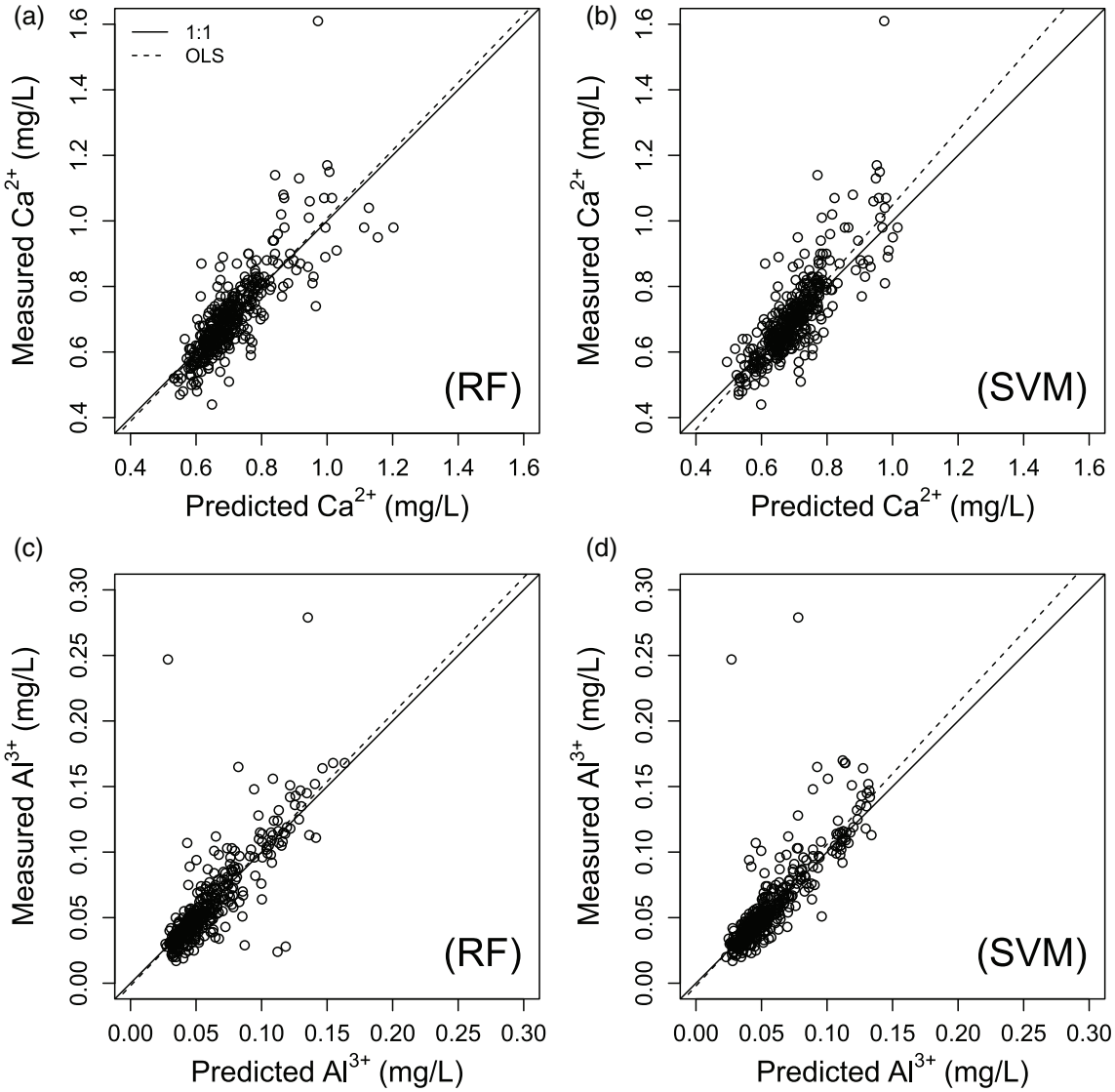

FIGURE 2 Comparison of measured calcium (Ca^{2+}) concentrations to predicted values from (a) the random forest (RF) model and (b) the support vector machine (SVM). The same comparison for total monomeric aluminium (Al^{3+}) from the (c) RF model and (d) SVM. The solid line is the 1:1 line and the dashed line is the ordinary least squares (OLS) line. The OLS models and Nash-Sutcliffe efficiencies were (a) $y = 1.03x–0.02$ ($r^2 = 0.66$; NS = 0.67), (b) $y = 1.14x–0.09$ ($r^2 = 0.66$; NS = 0.65), (c) $y = 1.04x$ ($r^2 = 0.65$; NS = 0.66), and (d) $y = 1.08x$ ($r^2 = 0.64$; NS = 0.64)

are also least accurate at peak concentrations (Figure 4). Our predicted 15-min NO_3^- concentration data was trained from only 382 grab samples, but effectively predicted the high resolution NO_3^- concentration data derived from our nitrate sensor (NS = 0.68 and 0.64 for the SVM and RF models, respectively; Figure 5). Similar to the comparison with discrete samples, the predicted 15-min values underpredicted periods of high NO_3^- concentrations. This underprediction is particularly apparent during the high concentrations in the winter of 2013–2014.

The SVM and RF predicted solute concentrations showed little degradation of the NS efficiency when removing just one independent variable (Table 2). The lack of NS efficiency degradation suggested that there is similar information held in multiple variables. Overall, the RF models had much higher baseline NS values and much lower NS changes when one variable was removed. The relative importance of different dependent variables was similar in both the SVM and RF models; FDOM, pH, and SC emerged as the most important variables. The largest individual decrease in NS was 0.185, which occurred when FDOM was removed from the SVM K^+ model, which began at 0.659. The largest decrease among the RF models was when FDOM was removed from the K^+ model; however, the decrease in NS was 0.046 (from 0.921). For both the RF and SVM models, other notable substantial decreases in NS were apparent for Ca^{2+} and Mg^{2+} when SC was removed, DOC when FDOM was removed, and NO_3^- when SC was removed. The least important independent variable was turbidity. There were some cases where the NS efficiency did not change or slightly increased with removal of an independent variable, such as with turbidity, SC, and DOY.

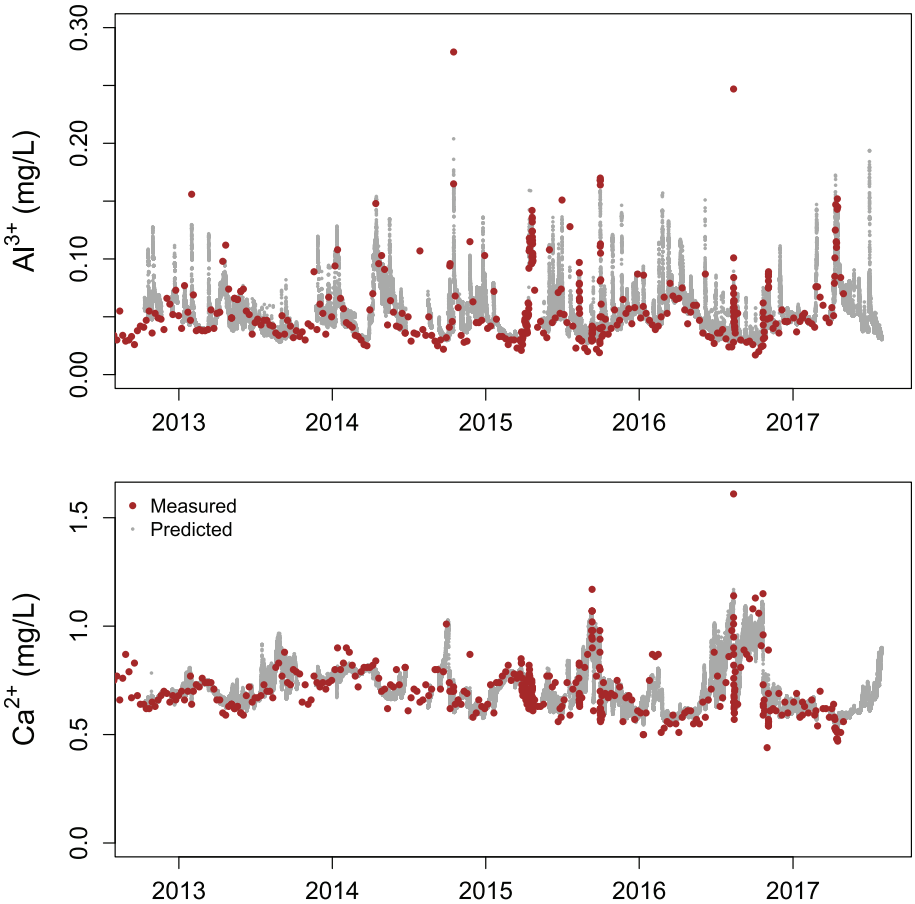

FIGURE 3 Time series of measured stream water samples (brown dots) and the predicted high-frequency predictions (grey points) for total monomeric aluminium (Al^{3+}) and calcium (Ca^{2+}) from the random forest model

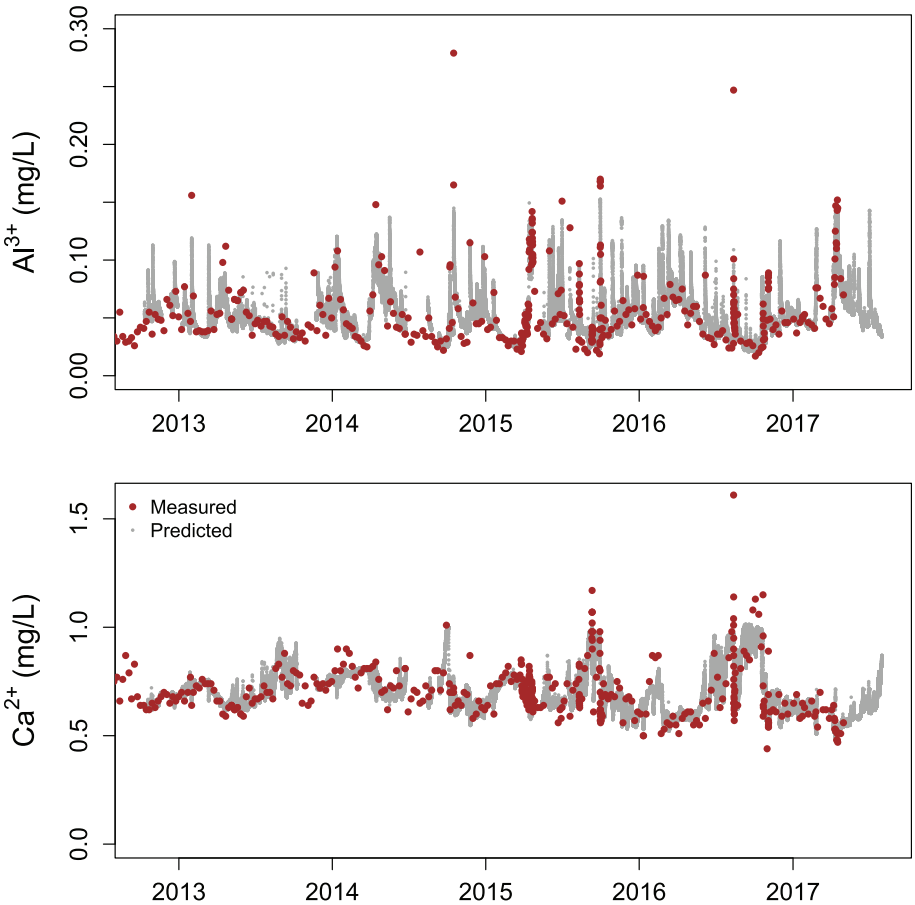

FIGURE 4 Time series of measured stream water samples (brown dots) and the predicted high-frequency predictions (grey points) for total monomeric aluminium (Al^{3+}) and calcium (Ca^{2+}) from the support vector machine model

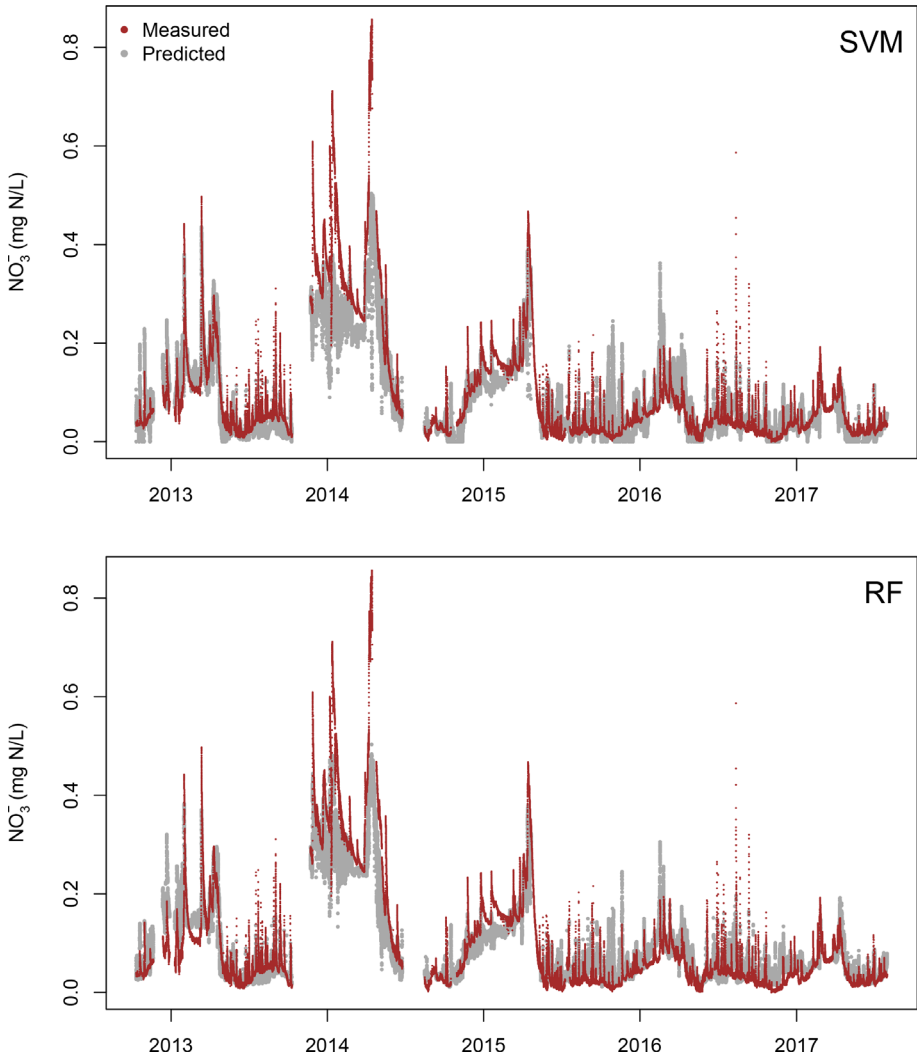

FIGURE 5 The time series of sensed and predicted high-frequency stream nitrate concentrations show similar variation in the SVM predictions (NS = 0.68) and RF predictions (NS = 0.64). The predicted concentrations (n = 146 213) are from models trained to predict nitrate in 382 discrete water samples. RF, random forest; SVM, support vector machine

TABLE 2 Change in the Nash-Sutcliffe (NS) efficiency of solute prediction models when one independent variable is removed

Variable removed	Al^{3+}	Ca^{2+}	Mg^{2+}	Na^+	K^+	SO_4^{2-}	Cl^-	NO_3^-	DOC	Si	Mean change
Support vector machine											
Baseline	0.747	0.74	0.853	0.836	0.659	0.807	0.471	0.882	0.681	0.716	
FDOM	−0.05	−0.016	−0.021	−0.031	−0.185	−0.039	−0.082	−0.012	−0.133	−0.015	−0.058
Turb	0	+0.002	−0.003	−0.017	0	−0.016	−0.05	−0.009	−0.021	−0.019	−0.013
pH	−0.077	−0.018	−0.016	−0.011	−0.028	−0.056	−0.024	−0.193	−0.009	−0.018	−0.045
SC	+0.001	−0.109	−0.107	−0.037	−0.052	0.001	−0.069	−0.073	+0.003	−0.06	−0.050
DO	−0.013	−0.017	−0.004	−0.029	−0.038	−0.016	−0.053	−0.101	−0.046	−0.027	−0.034
DOY	−0.012	−0.014	−0.018	−0.017	−0.056	0	−0.049	−0.017	0	−0.019	−0.020
Random forest											
Baseline	0.912	0.928	0.958	0.954	0.921	0.952	0.883	0.958	0.898	0.946	
FDOM	−0.021	−0.012	−0.007	−0.009	−0.046	−0.023	−0.019	−0.013	−0.029	−0.01	−0.019
Turb	−0.015	−0.014	−0.009	−0.008	−0.015	−0.016	−0.016	−0.013	−0.014	−0.014	−0.013
pH	−0.021	−0.017	−0.015	−0.01	−0.029	−0.022	−0.028	−0.029	−0.013	−0.019	−0.020
SC	−0.001	−0.028	−0.03	−0.011	−0.027	−0.012	−0.026	−0.032	−0.004	−0.017	−0.019
DO	−0.001	−0.014	−0.015	−0.007	−0.024	−0.01	−0.024	−0.018	−0.01	−0.008	−0.013
DOY	−0.002	−0.012	−0.012	−0.011	−0.044	−0.009	−0.031	−0.011	−0.011	−0.017	−0.016

Note: The baseline NS for the model with no variables removed (and all data points used; no training/testing sets) is shown for reference. The sensitivity of the solute prediction to each independent variable is indicated by a decrease in the NS efficiency relative to the model with no variable removed.

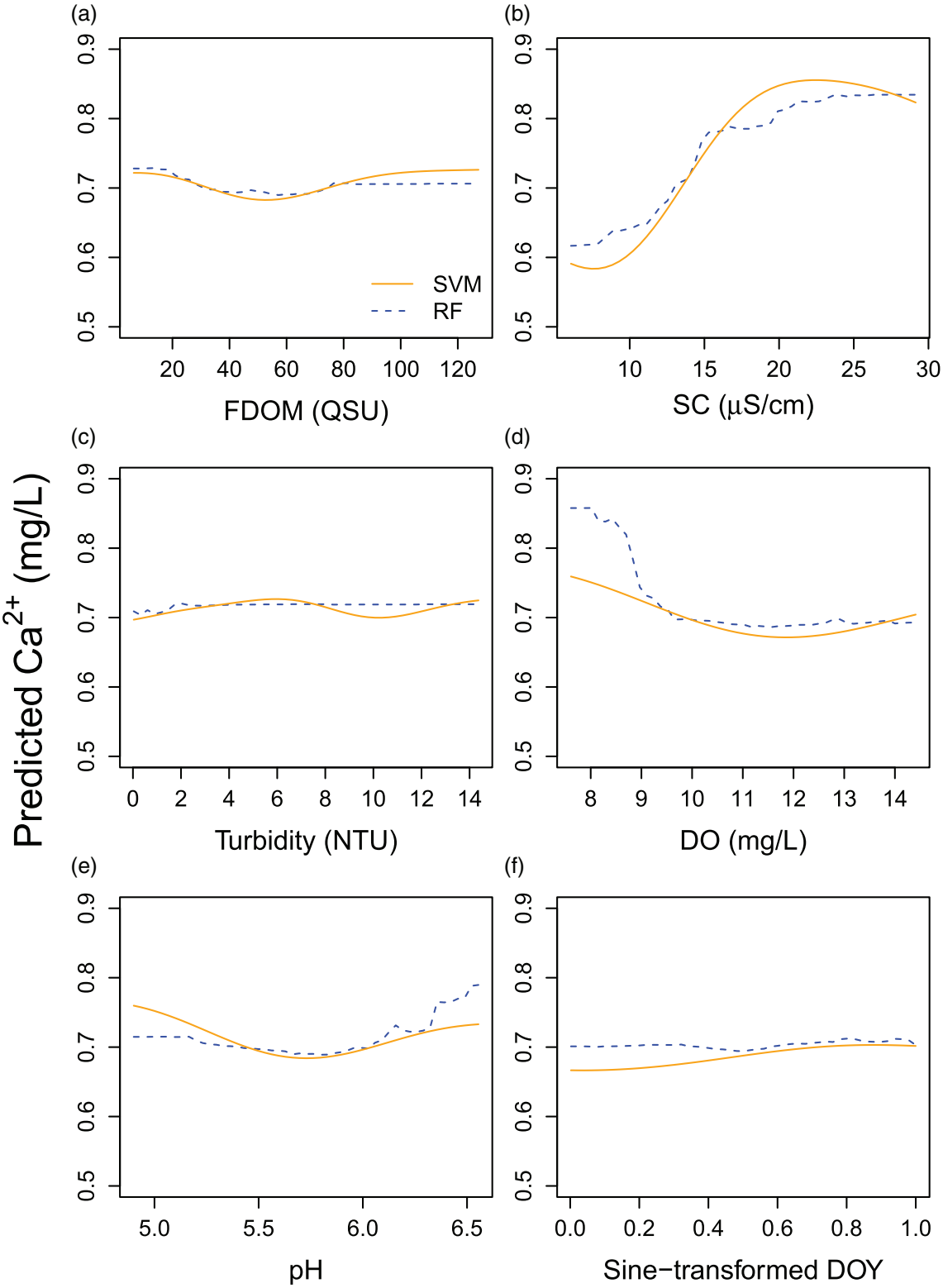

FIGURE 6 Partial dependence plots for the stream water calcium concentration RF and SVM models. The partial dependence on (a) fluorescent dissolved organic matter, (b) specific electrical conductivity, (c) turbidity, (d) dissolved oxygen, (e) pH, and (f) sine-transformed day of the year are shown. RF, random forest; SVM, support vector machine

Partial dependence plots for the Ca^{2+} and Al^{3+} highlighted the most important variables and the non-linear relationships between the independent variables and the solute concentrations (Figures 6 and 7). For Ca^{2+}, the largest changes in Ca^{2+} concentration are as SC changes from 10 to 20 µS/cm in both the SVM and RF models (Figure 6(b)). The RF model also shows higher Ca^{2+}

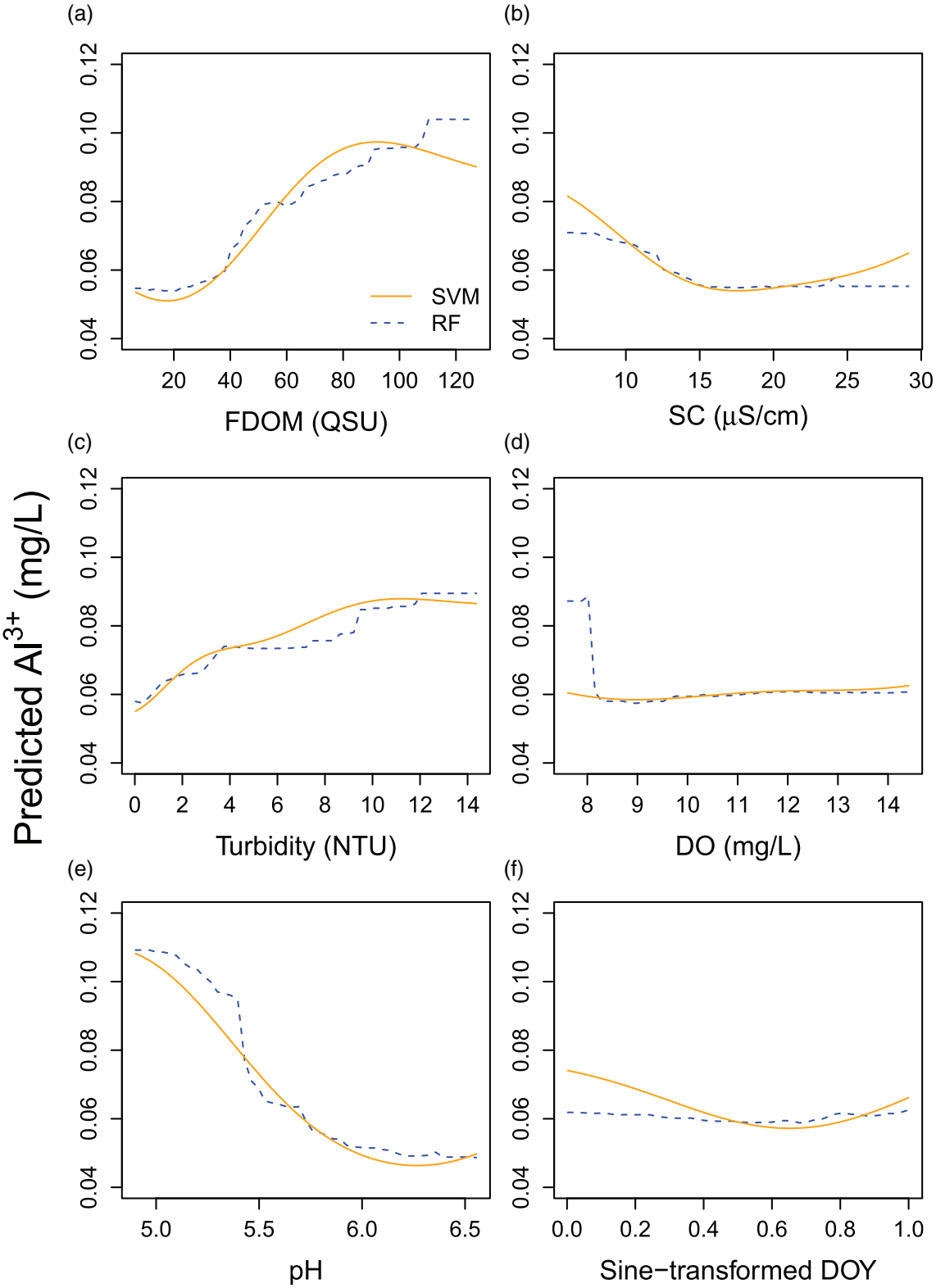

FIGURE 7 Partial dependence plots for the stream water total monomeric aluminium concentration RF and SVM models. The partial dependence on (a) fluorescent dissolved organic matter, (b) specific electrical conductivity, (c) turbidity, (d) dissolved oxygen, (e) pH, and (f) sine-transformed day of the year are shown. RF, random forest; SVM, support vector machine

concentration at low DO concentration and high pH (Figure 6(d) and (e)). The Al^{3+} predictions were dependent on more variables, with high Al^{3+} concentrations associated with high FDOM, high turbidity, and low pH (Figure 7). The SVM and RF partial dependence plots were generally similar for each independent variable, with the only notable divergence in the DO plot for RF where the

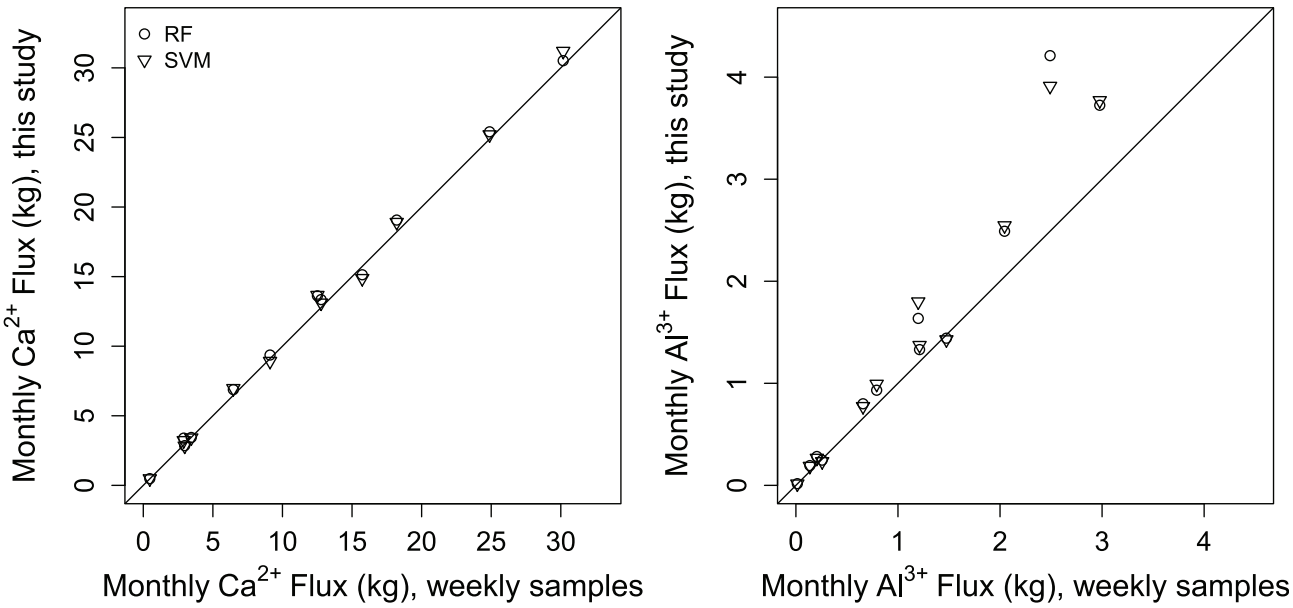

FIGURE 8 Comparison of stream calcium and monomeric aluminium monthly flux during 2016 using the weekly samples and the predicted high-frequency solute concentrations from this study. The 1:1 line is shown for reference

lowest DO values show a step increase in Al^{3+} concentration (Figure 7(d)).

Monthly fluxes of Ca^{2+} and Al^{3+} for 2016 using the predicted 15-min values were compared to the fluxes estimated with weekly discrete samples that were analysed in the laboratory. The monthly Ca^{2+} fluxes using the predicted 15-min values were 5% lower to 17% higher than estimates derived from period-weighted estimates derived from weekly samples, while the monthly Al^{3+} fluxes using the 15-min data were 9% lower to 69% higher than the estimates using weekly samples, with the largest differences during months when fluxes were high (Figure 8).

4 | DISCUSSION

We demonstrated that machine learning algorithms were effective at predicting the observed concentrations of major ions in Hubbard Brook stream water. Having access to such an extensive discrete water sampling and stream sensing data set allowed us to test the effectiveness of machine learning algorithms at predicting multiple non-sensed solutes at much higher frequency than previous studies. Our results demonstrate the potential to produce accurate, high-frequency solute predictions that can be used to gain new insights into hydrological and biogeochemical processes in catchments. Machine learning algorithms proficiently predict multiple Earth system variables (Lary et al., 2016; Olden et al., 2008; Raghavendra & Deka, 2014), thus it is not surprising that our solute predictions are generally accurate. The large number of discrete stream samples allowed the models to be well trained. Better predictions of stream solutes in the future could be gained from collection of a greater number of samples with high concentrations, particularly during storm flows, to better characterize the physio-chemical conditions when they occur. The optimal timing of discrete sampling could be identified by testing the effect on training machine learning algorithms. Further, a broader set of independent variables could be considered. For example, solute production and transport is influenced by meteorological conditions (e.g., Aulenbach, 2020; Wen et al., 2020) and thus the inclusion of meteorological variables such as air temperature, precipitation, vapour pressure deficit, snowpack condition, or potential evapotranspiration may result in better predictions.

While the development of new sensors may improve our understanding of high-frequency solute dynamics, our results suggest that machine learning applied to existing sensor data can provide accurate information about the temporal variability in concentrations of solutes for which no effective sensor has been developed. However, this method requires a robust sensing programme coupled with temporally and spatially intensive grab sampling to build models that are specific to each stream site of interest across a range of conditions. A robust set of discrete samples is vital to effective training of a machine learning algorithm. Once a significant number of models are built to represent a range of catchments, their transferability across places and time can be tested, but this is not yet possible due to a lack of data. As more paired stream solute and water quality sensing data sets are made available, it will be possible to assess whether solute prediction models like we have presented are stable when applied in other places or during long-term deployments. If these models are transferable, the cost of water quality monitoring could be greatly reduced and high temporal resolution information could be broadly generated.

The differences in predictive strength across solutes appeared to be related to their sources and biogeochemical dynamics. Generally, the solutes with mineral weathering sources least affected by ecosystem dynamics during this study period (Na^{+}, Ca^{2+}, Si, Mg^{2+}) or

relatively constant atmospheric deposition inputs (NO_3^-, SO_4^{2-}) were the best predicted. In contrast, those with either episodic or unpredictable atmospheric inputs (Cl^-) or those with potentially strong biotic as well as abiotic control (e.g., K^+; McDowell & Asbury, 1994) were much harder to predict. Both Cl^- and K^+ had low concentrations and were small contributors to the ionic charge balance. Among the major anions Cl^-, SO_4^{2-}, and NO_3^-, Cl^- was 16% of the average total microequivalents. Among the cations Ca^{2+}, Na^+, K^+, Mg^{2+}, and Al^{3+}, K^+ constituted 4% of the average total microequivalents. The contribution to the ionic charge balance was not always a factor in the predictability of a solute, as demonstrated by NO_3^- and Al^{3+} which contributed only 9 and 6% respectively, on average. In these cases, associations with other physiochemical variables – pH in particular – made them more predictable. Hubbard Brook streams are acidic and dilute (Likens & Buso, 2012), which was likely a major factor in our results, highlighting the need for similar studies in streams with different characteristics. It would be valuable to explore whether the approach we present here is effective in catchments with higher solute concentrations.

The sensitivity analysis uncovered the most important independent variables generally and for specific solutes (Table 2). Specific electrical conductivity (SC) was the most important independent variable for major ions, particularly the divalent cations Ca^{2+} and Mg^{2+}, likely due to the direct dependence of SC on the ionic strength of a solution (e.g., Miller et al., 1988). Model predictions were also sensitive to inclusion of FDOM and pH. FDOM is an effective proxy for dissolved DOC (e.g., Snyder et al., 2018; Wymore et al., 2018), which our analysis also suggested. However, K^+ was similarly sensitive to FDOM, which is not intuitive. It is possible that these ions may be associated with dissolved organic matter in our catchment, resulting in strong predictive power, but the nature of this association is not well-established. The strong dependence of Al^{3+}on pH was consistent with current understanding that aluminium biogeochemistry is acid-sensitive (Driscoll et al., 1980). In particular, the steep shift in the partial dependence of Al^{3+} on pH at pH = 5.5 in the RF model is interesting because this is a pH below which Al speciation shifts to Al^{3+} (Figure 7(e); Driscoll & Schecher, 1990). Predictions of NO_3^- were also sensitive to the inclusion of pH, perhaps because most stream water nitrate at Hubbard Brook is produced by nitrification (Pardo et al., 2004), which is an acidifying process. The visible partial dependence of high Ca^{2+} and Al^{3+} concentrations on very low DO was unexpected and may be related to transport of low DO groundwater (Figures 6(d) and 7(d); Krause et al., 2013). The low importance of turbidity to prediction of any solute is not surprising because it is a common proxy for suspended materials rather than dissolved solutes (Nasrabadi et al., 2016; Rügner et al., 2014). Turbidity is likely to be the most important available predictor of P fluxes in our watershed where nearly all P is exported in particulate forms (Meyer & Likens, 1979).

Predicted high-frequency solute concentrations have multiple applications, including source water tracing, catchment-scale mineral weathering rate estimation, water quality monitoring, and aquatic habitat assessment. Geogenic solutes provide useful tracers of stream water sources (e.g., Benettin et al., 2015; Burns et al., 2001); being able to trace water sources at higher temporal resolution will shed new light on streamflow generation processes. The higher Al^{3+} flux rates derived from our predicted concentrations may aid interpretations of mineral weathering rates, soil development, and soil responses to human-caused disturbances (Bailey, 2020; Johnson et al., 2000). Further, high-frequency streamflow tracing can help identify sources of biogeochemically dynamic solutes like nitrate (Pardo et al., 2020). Or, when paired with other high-frequency sensed variables such as dissolved oxygen or CO_2, stream solute dynamics can be linked with ecological phenomena such as stream metabolism. Further application of high-frequency water quality predictions would improve aquatic condition assessments because it may allow for exploration of exposure to high pollutant concentrations that are often episodic and challenging to measure and predict. For example, free ionic Al^{3+} is toxic to gilled organisms (e.g., Kroglund & Finstad, 2003), and predictions of episodic Al^{3+} increases in concentration would allow quantification of organism exposure to toxic concentrations.

5 | CONCLUSION

A wide range of stream solutes was accurately predicted at 15-min intervals using two machine learning algorithms driven by data from an array of physico-chemical properties derived from aquatic sensors and extensive discrete sampling across a range of flow conditions. More extensive data sets or other machine learning algorithms will likely improve the accuracy of predicted stream solutes for which no known physical or chemical proxy is available with current sensor technology, thus opening new insights into the high-frequency variability of non-sensed solutes. Such information will allow improved estimation of stream water solute concentrations and exports, especially those solutes that increase in concentration with discharge.

ACKNOWLEDGEMENTS

The authors thank Tammy Wooster who led the stream sampling, Jeff Merrriam who led the laboratory analysis, Brenda Minicucci who maintained the stream sample data, Lisle Snyder who maintained the sensors, and Jody Potter who maintained the sensor data. The authors also thank two anonymous reviewers for helpful comments on an earlier version of this manuscript.

DATA AVAILABILITY STATEMENT

The discrete water chemistry and stream discharge data that support the findings of this study are openly available in the Environmental Data Initiative at http://doi.org/doi:10.6073/pasta/87584eda806dd5a480423b6bfefec577 and http://doi.org/doi:10.6073/pasta/4022d829f3a1fa4057b63b5db8b1a172. The water quality sensing data that support the findings of this study are openly available from the HydroShare at https://www.hydroshare.org/resource/8217eab0997d493782ff321ca5f95f28/.

ORCID

Mark B. Green https://orcid.org/0000-0002-7415-7209
Scott W. Bailey https://orcid.org/0000-0002-9160-156X
John L. Campbell https://orcid.org/0000-0003-4956-1696
William H. McDowell https://orcid.org/0000-0002-8739-9047

REFERENCES

Allan, J. D., & Ibañez Castillo, M. M. (2009). *Stream ecology: Structure and function of running waters*. Dordrecht: Springer.

Aubert, A. H., & Breuer, L. (2016). New seasonal shift in in-stream diurnal nitrate cycles identified by mining high-frequency data. *PLoS One*, *11*, e0153138.

Aulenbach, B. T. (2020). Effects of climate-related variability in storage on streamwater solute concentrations and fluxes in a small forested watershed in the southeastern United States. *Hydrological Processes*, *34*(2), 189–208.

Aulenbach, B. T., Burns, D. A., Shanley, J. B., Yanai, R. D., Bae, K., Wild, A. D., Yang, Y., & Yi, D. (2016). Approaches to stream solute load estimation for solutes with varying dynamics from five diverse small watersheds. *Ecosphere*, *7*, e01298.

Bailey, A. S., Hornbeck, J. W., Campbell, J. L., & Eagar, C. (2003). *Hydrometeorological database for Hubbard Brook Experimental Forest: 1955–2000. U.S* (p. 305). Forest Service.

Bailey, S. W. (2020). *Tracking the fate of plagioclase weathering products. Page biogeochemical cycles: Ecological drivers and environmental impact*. American Geophysical Union.

Bailey, S. W., Brousseau, P. A., McGuire, K. J., & Ross, D. S. (2014). Influence of landscape position and transient water table on soil development and carbon distribution in a steep, headwater catchment. *Geoderma*, *226–227*, 279–289.

Bailey, S. W., Buso, D. C., & Likens, G. E. (2003). Implications of sodium mass balance for interpreting the calcium cycle of a forested ecosystem. *Ecology*, *84*, 471–484.

Bailey, S. W., McGuire, K. J., Ross, D. S., Green, M. B., & Fraser, O. L. (2019). Mineral weathering and Podzolization control acid neutralization and streamwater chemistry gradients in upland glaciated catchments, northeastern United States. *Frontiers in Earth Science*, *7*, 63.

Basu, N. B., Destouni, G., Jawitz, J. W., Thompson, S. E., Loukinova, N. V., Darracq, A., Zanardo, S., Yaeger, M., Sivapalan, M., Rinaldo, A., & Rao, P. S. C. (2010). Nutrient loads exported from managed catchments reveal emergent biogeochemical stationarity. *Geophysical Research Letters*, *37*(23), L23404.

Baudron, P., Alonso-Sarría, F., García-Aróstegui, J. L., Cánovas-García, F., Martínez-Vicente, D., & Moreno-Brotóns, J. (2013). Identifying the origin of groundwater samples in a multi-layer aquifer system with random forest classification. *Journal of Hydrology*, *499*, 303–315.

Benettin, P., Bailey, S. W., Campbell, J. L., Green, M. B., Rinaldo, A., Likens, G. E., McGuire, K. J., & Botter, G. (2015). Linking water age and solute dynamics in streamflow at the Hubbard Brook Experimental Forest, NH, USA. *Water Resources Research*, *51*, 9256–9272.

Bernhardt, E. S., Likens, G. E., Buso, D. C., & Driscoll, C. T. (2003). In-stream uptake dampens effects of major forest disturbance on watershed nitrogen export. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 10304–10308.

Bettez, N. D., Duncan, J. M., Groffman, P. M., Band, L. E., O'Neil-Dunne, J., Kaushal, S. S., Belt, K. T., & Law, N. (2015). Climate variation overwhelms efforts to reduce nitrogen delivery to coastal waters. *Ecosystems*, *18*, 1319–1331.

Bormann, F. H., & Likens, G. E. (1979). *Pattern and process in a forested ecosystem: Disturbance, development, and the steady state based on the Hubbard brook ecosystem study*. Springer-Verlag.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Burns, D. A., McDonnell, J. J., Hooper, R. P., Peters, N. E., Freer, J. E., Kendall, C., & Beven, K. (2001). Quantifying contributions to storm runoff through end-member mixing analysis and hydrologic measurements at the Panola Mountain research watershed (Georgia, USA). *Hydrological Processes*, *15*, 1903–1924.

Dittman, J. A., Shanley, J. B., Driscoll, C. T., Aiken, G. R., Chalmers, A. T., & Towse, J. E. (2009). Ultraviolet absorbance as a proxy for total dissolved mercury in streams. *Environmental Pollution*, *157*, 1953–1956.

Driscoll, C. T., Baker, J. P., Bisogni, J. J., & Schofield, C. L. (1980). Effect of aluminium speciation on fish in dilute acidified waters. *Nature*, *284*, 161–164.

Driscoll, C. T., & Schecher, W. D. (1990). The chemistry of aluminum in the environment. *Environmental Geochemistry and Health*, *12*(1–2), 28–49.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. J., & Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems*, *9*, 155–161.

Eshleman, K. N., & Sabo, R. D. (2016). Declining nitrate-N yields in the Upper Potomac River Basin: What is really driving progress under the Chesapeake Bay restoration? *Atmospheric Environment*, *146*, 280–289.

Gillin, C. P., Bailey, S. W., McGuire, K. J., & Gannon, J. P. (2015). Mapping of hydropedologic spatial patterns in a steep headwater catchment. *Soil Science Society of America Journal*, *79*, 440–453.

Godsey, S. E., Kirchner, J. W., & Clow, D. W. (2009). Concentration-discharge relationships reflect chemostatic characteristics of US catchments. *Hydrological Processes*, *23*, 1844–1864.

Greenwell, B. M. (2017). pdp: An R package for constructing partial dependence plots. *The R Journal*, *9*, 421–436.

Inserillo, E. A., Green, M. B., Shanley, J. B., & Boyer, J. N. (2017). Comparing catchment hydrologic response to a regional storm using specific conductivity sensors. *Hydrological Processes*, *31*, 1074–1085.

Johnson, C. E., Driscoll, C. T., Siccama, T. G., & Likens, G. E. (2000). Element fluxes and landscape position in a northern hardwood forest watershed ecosystem. *Ecosystems*, *3*, 159–184.

Johnson, N. M., Likens, G. E., Bormann, F. H., Fisher, D. W., & Pierce, R. S. (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.

Kim, Y., Johnson, M. S., Knox, S. H., Black, T. A., Dalmagro, H. J., Kang, M., Kim, J., & Baldocchi, D. (2020). Gap-filling approaches for eddy covariance methane fluxes: A comparison of three machine learning algorithms and a traditional method with principal component analysis. *Global Change Biology*, *26*, 1499–1518.

Kisi, O., & Parmar, K. S. (2016). Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water pollution. *Journal of Hydrology*, *534*, 104–112.

Koenig, L. E., Shattuck, M. D., Snyder, L. E., Potter, J. D., & McDowell, W. H. (2017). Deconstructing the effects of flow on DOC, nitrate, and major ion interactions using a high-frequency aquatic sensor network. *Water Resources Research*, *53*, 10655–10673.

Kottek, M., Grieser, J., Beck, C., Rudolf, B., & Rubel, F. (2006). World map of the Köppen-Geiger climate classification updated. *Meteorologische Zeitschrift*, *15*, 259–263.

Krause, S., Tecklenburg, C., Munz, M., & Naden, E. (2013). Streambed nitrogen cycling beyond the hyporheic zone: Flow controls on horizontal patterns and depth distribution of nitrate and dissolved oxygen in the upwelling groundwater of a lowland river. *Journal of Geophysical Research: Biogeosciences*, *118*, 54–67.

Kroglund, F., & Finstad, B. (2003). Low concentrations of inorganic monomeric aluminum impair physiological status and marine survival of Atlantic salmon. *Aquaculture*, *222*, 119–133.

Lannergård, E. E., Ledesma, J. L. J., Fölster, J., & Futter, M. N. (2019). An evaluation of high frequency turbidity as a proxy for riverine total phosphorus concentrations. *Science of the Total Environment*, *651*, 103–113.

Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L. (2016). Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, *7*, 3–10.

Lawrence, G. B., Driscoll, C. T., & Fuller, R. D. (1988). Hydrologic control of aluminum chemistry in an acidic headwater stream. *Water Resources Research*, *24*, 659–669.

Lawrence, G. B., Hazlett, P. W., Fernandez, I. J., Ouimet, R., Bailey, S. W., Shortle, W. C., Smith, K. T., & Antidormi, M. R. (2015). Declining acidic deposition begins reversal of forest-soil acidification in the northeastern U.S. and eastern Canada. *Environmental Science & Technology*, *49*, 13103–13111.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*, 6.

Likens, G. E. (2013). *Biogeochemistry of a forested ecosystem* (3rd ed.). New York, NY: Springer.

Likens, G. E., & Buso, D. C. (2012). Dilution and the elusive baseline. *Environmental Science & Technology*, *46*, 4382–4387.

Likens, G. E., Driscoll, C. T., & Buso, D. C. (1996). Long-term effects of acid rain: Response and recovery of a forest ecosystem. *Science*, *272*, 244–246.

McDowell, W. H., & Asbury, C. E. (1994). Export of carbon, nitrogen, and major ions from three tropical montane watersheds. *Limnology and Oceanography*, *39*, 111–125.

Mewes, B., Oppel, H., Marx, V., & Hartmann, A. (2020). Information-based machine learning for tracer signature prediction in Karstic environments. *Water Resources Research*, *56*, e2018WR024558.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2019). *e1071: Misc functions of the department of statistics, probability theory group (formerly: E1071)*. Wien.

Meyer, J. L., & Likens, G. E. (1979). Transport and transformation of phosphorus in a forest stream ecosystem. *Ecology*, *60*(6), 1255–1269.

Miller, R., Bradford, W., & Peters, N. (1988). *Specific conductance; theoretical considerations and application to analytical quality control* (p. 2311). United States Geological Survey.

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of Hydrology*, *10*, 282–290.

Nasrabadi, T., Ruegner, H., Sirdari, Z. Z., Schwientek, M., & Grathwohl, P. (2016). Using total suspended solids (TSS) and turbidity as proxies for evaluation of metal transport in river water. *Applied Geochemistry*, *68*, 1–9.

Neal, C., Reynolds, B., Kirchner, J. W., Rowland, P., Norris, D., Sleep, D., Lawlor, A., Woods, C., Thacker, S., Guyatt, H., & Vincent, C. (2013). High-frequency precipitation and stream water quality time series from Plynlimon, Wales: An openly accessible data resource spanning the periodic table. *Hydrological Processes*, *17*(2013), 2531–2539.

Olden, J. D., Lawler, J. J., & Poff, N. L. (2008). Machine learning methods without tears: A primer for ecologists. *The Quarterly Review of Biology*, *83*, 171–193.

Pardo, L. H., Green, M. B., Bailey, S. W., McGuire, K. J., & McDowell, W. H. (2020). Identifying controls on nitrate sources and flowpaths in a forested catchment using a hydropedological framework.

Pardo, L. H., Kendall, C., Pett-Ridge, J., & Chang, C. C. Y. (2004). Evaluating the source of streamwater nitrate using δ15N and δ18O in nitrate in two watersheds in New Hampshire, USA. *Hydrological Processes*, *18*, 2699–2712.

Pellerin, B. A., Downing, B. D., Kendall, C., Dahlgren, R. A., Kraus, T. A., Saracenco, J., Spencer, R. G. M., & Bergamaschi, B. A. (2009). Assessing the sources and magnitude of diurnal nitrate variability in the San Joaquin River (California) with an in situ optical nitrate sensor and dual nitrate isotopes. *Freshwater Biology*, *54*, 376–387.

Raghavendra, N. S., & Deka, P. C. (2014). Support vector machine applications in the field of hydrology: A review. *Applied Soft Computing*, *19*, 372–386.

Ross, M. R., Nippgen, F., Hassett, B. A., McGlynn, B. L., & Bernhardt, E. S. (2018). Pyrite oxidation drives exceptionally high weathering rates and geologic CO_2 release in mountaintop-mined landscapes. *Global Biogeochemical Cycles*, *32*(8), 1182–1194.

Rügner, H., Schwientek, M., Egner, M., & Grathwohl, P. (2014). Monitoring of event-based mobilization of hydrophobic pollutants in rivers: Calibration of turbidity as a proxy for particle facilitated transport in field and laboratory. *Science of the Total Environment*, *490*, 191–198.

Schleppi, P., Waldner, P. A., & Stähli, M. (2006). Errors of flux integration methods for solutes in grab samples of runoff water, as compared to flow-proportional sampling. *Journal of Hydrology*, *319*, 266–281.

Sebestyen, S. D., Dorrance, C., Olson, D. M., Verry, E. S., Kolka, R. K., Elling, A. E., & Kyllander, R. (2011). Long-term monitoring sites and trends at the Marcell Experimental Forest. In R. K. Kolka, S. D. Sebestyen, E. S. Verry, & K. N. Brooks (Eds.), *Peatland biogeochemistry and watershed hydrology at the Marcell Experimental Forest* (pp. 15–71). Boca Raton, FL: CRC Press.

Siccama, T. G., Fahey, T. J., Johnson, C. E., Sherry, T. W., Denny, E. G., Girdler, E. B., Likens, G. E., & Schwarz, P. A. (2007). Population and biomass dynamics of trees in a northern hardwood forest at Hubbard Brook. *Canadian Journal of Forest Research*, *37*, 737–749.

Snyder, L., Potter, J. D., & McDowell, W. H. (2018). An evaluation of nitrate, fDOM, and turbidity sensors in New Hampshire streams. *Water Resources Research*, *54*, 2466–2479.

Swistock, B. R., Edwards, P. J., Wood, F., & Dewalle, D. R. (1997). Comparison of methods for calculating annual solute exports from six forested Appalachian watersheds. *Hydrological Processes*, *11*, 655–669.

Tan, G., Yan, J., Gao, C., & Yang, S. (2012). Prediction of water quality time series data based on least squares support vector machine. *Procedia Engineering*, *31*, 1194–1199.

USDA Forest Service, Northern Research Station. (2019). USFS Durham lab water analysis method detection limit (MDL) limit of quantification (LOQ), 2010 - present ver 1. *Environmental Data Initiative*. https://doi.org/10.6073/pasta/890b1fadb61d3e86dc6f3d9afea79705.

Valett, H. M., Crenshaw, C. L., & Wagner, P. F. (2002). Stream nutrient uptake, forest succession, and biogeochemical theory. *Ecology*, *83*, 2888–2901.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer Science & Business Media.

Wen, H., Perdrial, J., Abbott, B. W., Bernal, S., Dupas, R., Godsey, S. E., Harpold, A., Rizzo, D., Underwood, K., Adler, T., & Sterle, G. (2020). Temperature controls production but hydrology regulates export of dissolved organic carbon at the catchment scale. *Hydrology and Earth System Sciences*, *24*, 945–966.

Wymore, A. S., Potter, J., Rodríguez-Cardona, B., & McDowell, W. H. (2018). Using in-situ optical sensors to understand the biogeochemistry of dissolved organic matter across a stream network. *Water Resources Research*, *54*, 2949–2958.

How to cite this article: Green MB, Pardo LH, Bailey SW, et al. Predicting high-frequency variation in stream solute concentrations with water quality sensors and machine learning. *Hydrological Processes*. 2021;35:e14000. https://doi.org/10.1002/hyp.14000