Environmental Science Processes & Impacts

Open Access Article. Published on 03 May 2016. Downloaded on 11/18/2025 5:11:36 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

PAPER

View Article Online
View Journal | View Issue

Cite this: *Environ. Sci.: Processes Impacts*, 2016, **18**, 590

Statistical generation of training sets for measuring NO_3^-, NH_4^+ and major ions in natural waters using an ion selective electrode array†

Amy V. Mueller‡*[a] and Harold F. Hemond[b]

Knowledge of ionic concentrations in natural waters is essential to understand watershed processes. Inorganic nitrogen, in the form of nitrate and ammonium ions, is a key nutrient as well as a participant in redox, acid–base, and photochemical processes of natural waters, leading to spatiotemporal patterns of ion concentrations at scales as small as meters or hours. Current options for measurement *in situ* are costly, relying primarily on instruments adapted from laboratory methods (*e.g.*, colorimetric, UV absorption); free-standing and inexpensive ISE sensors for NO_3^- and NH_4^+ could be attractive alternatives if interferences from other constituents were overcome. Multi-sensor arrays, coupled with appropriate non-linear signal processing, offer promise in this capacity but have not yet successfully achieved signal separation for NO_3^- and NH_4^+ *in situ* at naturally occurring levels in unprocessed water samples. A novel signal processor, underpinned by an appropriate sensor array, is proposed that overcomes previous limitations by explicitly integrating basic chemical constraints (*e.g.*, charge balance). This work further presents a rationalized process for the development of such *in situ* instrumentation for NO_3^- and NH_4^+, including a statistical-modeling strategy for instrument design, training/calibration, and validation. Statistical analysis reveals that historical concentrations of major ionic constituents in natural waters across New England strongly covary and are multi-modal. This informs the design of a statistically appropriate training set, suggesting that the strong covariance of constituents across environmental samples can be exploited through appropriate signal processing mechanisms to further improve estimates of minor constituents. Two artificial neural network architectures, one expanded to incorporate knowledge of basic chemical constraints, were tested to process outputs of a multi-sensor array, trained using datasets of varying degrees of statistical representativeness to natural water samples. The accuracy of ANN results improves monotonically with the statistical representativeness of the training set (error decreases by ~5×), while the expanded neural network architecture contributes a further factor of 2–3.5 decrease in error when trained with the most representative sample set. Results using the most statistically accurate set of training samples (which retain environmentally relevant ion concentrations but avoid the potential interference of humic acids) demonstrated accurate, unbiased quantification of nitrate and ammonium at natural environmental levels (±20% down to <10 μM), as well as the major ions Na^+, K^+, Ca^{2+}, Mg^{2+}, Cl^-, and SO_4^{2-}, in unprocessed samples. These results show promise for the development of new *in situ* instrumentation for the support of scientific field work.

Received 27th January 2016
Accepted 19th April 2016

DOI: 10.1039/c6em00043f

rsc.li/process-impacts

Environmental impact

This work presents a rationalized process for *in situ* instrumentation development with a focus on the measurement of ammonium and nitrate, key players in important environment-human coupled processes (agricultural runoff and water treatment) as well as participants in redox, acid–base, photochemical, and biologically driven transformation pathways. Built on a statistical-modeling strategy for instrument design, training/calibration, and validation, the process suggests a novel methodology for overcoming signal interferences. The analysis presented informs our understanding of the (highly covarying) statistical relationship of ions in fresh waters, while the tested architecture enables *in situ* measurements in non-processed samples down to micromolar levels, serving to increase spatiotemporal resolution of natural studies, enable adaptive sampling, and optimize relevancy of the limited number of grab samples included in most campaigns.

[a]*Massachusetts Institute of Technology, Cambridge, MA, USA. E-mail: amym@alum.mit.edu*

[b]*Massachusetts Institute of Technology, 77 Massachusetts Ave. 48-425, Cambridge, MA, USA*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c6em00043f

‡ Present address: University of Washington School of Oceanography, Seattle, WA, 98195, USA.

This journal is © The Royal Society of Chemistry 2016

Open Access Article. Published on 03 May 2016. Downloaded on 11/18/2025 5:11:36 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

1 Introduction

In natural waters, biological productivity, floral and faunal composition, and suitability of waters for human use are profoundly influenced by the concentrations of major ions as well as ionic forms of key nutrients that occur at lesser concentrations. Ammonium and nitrate are of particular interest due to their varied roles, including contributions to coastal eutrophication[1–3] from fluxes attributed to a wide variety of non-point sources (*e.g.*, fertilizers, wastewater treatment, or land use change) as well as roles in acid–base, redox, and photochemistry. These species furthermore undergo biologically driven transformations that alter both the quantity and form of biologically available nitrogen (*i.e.*, nitrification, denitrification, uptake, DNRA, and anammox), resulting in water compositions that provide little insight into original nutrient sources or forms – knowledge needed for diagnosis or remediation – unless measurements are obtained at adequate temporal and spatial frequencies directly within the system of interest. However, ammonium and nitrate remain particularly difficult to measure *in situ* and at the appropriate spatiotemporal scales.

To date, most commercial field instrumentation for measuring N species has been based on wet chemistry and spectrophotometric methods (*e.g.*, EnviroTech NAS-3X/EcoLAB, YSI 9600, Systea NPA/DPA) which require on-board reagents, pumps, and waste containment and generally have significant power and maintenance requirements. UV absorption may be a viable alternative for the measurement of nitrate in marine environments[4] but is less applicable in fresh waters where humic acids can play a major role. Because of their small size and low power requirements, ion selective electrodes (ISEs) are attractive alternatives despite challenges such as cross-ion interferences and limitations in the lowest concentrations for which linear (Nernstian) responses are achieved. As examples of the latter, ISE-based instruments recently introduced by YSI (6820) and Hach (Hydrolab series) for the *in situ* measurement of NO_3^- and NH_3^+ (typically two- or three-sensor arrays also including a measurement of chloride and/or electrical conductivity) have reported uncertainty of the maximum of ±10% or 2 mg N per L.[5] This implies a best-case scenario detection limit of >143 μM for both analytes, whereas concentrations of tens of μM (or less) are of interest or concern in many natural waters.[6]

Commercial ISEs themselves have response limits in this range (*e.g.*, see Table SI.1†). The challenge, therefore, in lowering ISE detection limits without sample pre-processing (*i.e.*, for direct *in situ* use) is in both (1) utilization of the response in the non-Nernstian region near the detection limit and (2) overcoming interferences from other naturally occurring ions that are often present in natural waters at levels 10–1000 times that of ammonium or nitrate. Multi-sensor arrays, coupled with non-linear signal processing methods, present one promising strategy for doing so by simultaneously quantifying both target and interfering species and retrieving deconvolved data from the set of highly non-linear sensor responses. Such multi-ISE instruments, initially conceived for ion measurements in biological liquids,[7] have been tested for analysis of heavy metals,[8–10] inorganic pollutants,[10,11] and small sets of inorganic ions[12,13] in a range of environmental contexts (*e.g.*, simulated polluted groundwaters). Conventional non-linear signal processing methods, *e.g.*, partial least squares regression, artificial neural networks (ANNs), and more recently Bayesian (blind) source separation,[14,15] have been successfully used to reconstruct concentrations of interest from the suite of interference-laden sensor signals. Although nitrate and ammonium ions have been targeted in specific applications where concentrations are expected to be much higher than natural levels (*e.g.*, fertigation,[16,17] eutrophied surface waters[18]), measurements at lower concentrations inevitably show systematic bias from total salinity,[18] leading researchers to suggest the use of a more comprehensive set of ISEs.

Further, in previous work, representativeness (*i.e.*, composition similarity to waters targeted for study) of samples used to calibrate (or "train", in the case of ANNs) signal processing algorithms has been identified as a primary driving force of system quality.[16] In general, ANN calibration has been conducted using individual standards bracketing target concentrations or synthetic samples with a fixed background (approximating the mean of targeted waters) to which ions of interest are added at fixed increments. In limited cases, the background constituents have further been systematically varied at 2–3 concentrations across target ranges[10] or a limited number of field samples have been incorporated[16,18] into the calibration set. While all of these methods bracket the concentration ranges of interest, they (1) are unlikely to be statistically representative of actual environmental waters if all calibration samples are equally weighted when presented to the mathematical algorithms and (2) inherently fail to capture covariance among analytes, which theoretically may be exploited to improve performance. The use of large training or calibration sets of actual environmental water samples could mitigate both issues but presents further challenges: (1) the need for sample processing to avoid sample-changing biological activity during transport and storage (which itself alters samples from *in situ* conditions), (2) the risk of non-representativeness if collection of samples is restricted to a small geographic area or time interval, and (3) the presence of additional (and possibly interfering) compounds such as DOC or humic acids which one may wish to avoid in early stages of development.

This work investigates the hypothesis that the development of *in situ* instrumentation to support scientific field studies of nitrate and ammonium ions (and other ionic species) can be enabled and expedited by combining (1) an improved understanding of the statistical characteristics of the target waters and (2) a signal processor which is able to take advantage of information available in (a) statistical relationships among sample analytes, (b) interference-laden ISE responses, and (c) an *a priori* understanding of the chemistry governing all surface waters. A statistical model of ion concentrations in the target environment is developed and used to create a synthetic training set that explicitly honors the covarying statistics of ionic constituents. To test the proposed hypothesis, a standard

Open Access Article. Published on 03 May 2016. Downloaded on 11/18/2025 5:11:36 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

ANN and an expanded ANN incorporating *a priori* chemical knowledge of the system (*i.e.*, charge balance, conductivity)[19] are trained using this sample set as measured by a comprehensive sensor array, as well as two other training sets with decreasing statistical representativeness. Results are validated by measuring the accuracy with which ion concentrations are estimated in independent statistically representative synthetic environmental samples. Additional tables available as ESI† are referenced in the text.

2 Experimental (materials and methods)

2.1 Sensor hardware

The array of sensors, comprised primarily of ion selective electrodes but also including sensors measuring conductivity, temperature, and pH (parameters affecting the ISE response and characterizing system conditions), was selected to ensure a measurable response (on at least one but often several channels) to all ions making up the majority of charge balance (>95%) of natural waters, *i.e.*, Na^+, K^+, Cl^-, Ca^{2+}, Mg^{2+}, SO_4^{3-}, NH_4^+, NO_3^-, and the pH and carbonate systems. Commercial ISEs (details in Table SI.1†) were purchased to measure the following: Na^+ (glass), Na^+ (solid state), K^+, Cl^-, Ca^{2+}, hardness (Mg^{2+} and Ca^{2+}), NH_4^+, and NO_3^-. Sensors marketed for the measurement of CO_3^{2-} and SO_4^{3-} were also purchased and used for data collection in the array; however these data were ultimately discarded as the relevant sample concentration ranges were below the response limit of the sensors.

2.2 Non-linear signal processor: artificial neural networks (ANNs)

While a number of algorithms have demonstrated utility for processing data from ISE sensor arrays, ANNs were selected for this work because (1) no assumptions need to be made about the form of the sensor responses (*e.g.*, semi-empirical/physics-based descriptions such as the Nikolsky–Eisenman equation) and (2) the underlying mathematical structure is amenable to integration of *a priori* chemical knowledge of the system. A brief introduction of ANNs is provided here for context, while the reader is referred to the literature for further details of standard ANN algorithms[12,20–22] and the strategy developed by the authors to integrate chemical information into the ANN architecture.[19]

An ANN is an unconstrained non-linear function estimator modeled on a (conceptual) understanding of the human neural structure. The mathematical representation is a topology of interconnected neurons whose firing triggers (or fails to trigger) subsequent neurons based on the relative strength of the interconnections. The number of inputs need not match the number of outputs, and in fact ANNs are well-suited for solving over-constrained systems. A prototypical structure is given in Fig. SI.1† along with additional details of architecture, parameterization, and training methodologies. It is important to note that ANN problems have no closed form solution, *i.e.*, it is not possible *a priori* to predict the optimal parameterization or number of required training samples/iterations. Prior work has, however, investigated sensitivity to a choice of parameter values for chemical applications,[12,23–25] providing a starting point for analysis of new problems.[12,26] In spite of this, it is still typically necessary to explore the permutations of possible ANN parameters through trial-and-error to find an optimal system on an application-by-application basis as system results can be determined more strongly by parameterization than by training data.[27]

As described in ref. 19, the ANN architecture integrating chemical knowledge takes advantage of the built-in neuron signal architecture to create output neurons that calculate conductivity (a property-weighted sum of all ions) and charge balance (a charge-weighted sum of all ions). Such calculations are possible because the hardware described above measures conductivity as well as ions representing >95% of the charge balance of natural waters. Error in these signals is used in addition to error in the ion concentrations to drive system training.

2.3 Multidimensional probability density function for ions in surface waters

To create a statistically representative training set, it was first necessary to build a statistical model of the target waters. Developing this model required data for a large suite of water samples for which all ion concentrations (as stated above: Na^+, K^+, Cl^-, Ca^{2+}, Mg^{2+}, SO_4^{3-}, NH_4^+, NO_3^-, and the pH and carbonate systems) were accurately known; it was further desirable that such data cover a wide range of surface water characteristics to ensure the applicability of the resulting instrumentation to varied field conditions. New England was selected as the study area as (1) numerous historical datasets are available, *e.g.*, from USGS monitoring efforts, (2) waters were expected to vary from soft to hard and oligotrophic to eutrophic, and (3) future in-field validation efforts would be facilitated.

Statistical characterization of New England waters was based on 50 years of historical data, downloaded from the USGS database for water quality samples.[28] Between 25 000 and 65 000 data points (measurements of a single analyte at a given site and time) were downloaded for each of the five states (MA, CT, VT, NH, ME). While such data can be used directly to estimate underlying one-dimensional probability density functions (PDFs – in this case, scaled histograms) for each analyte, which are informative in terms of range and frequency of particular ion concentrations, such PDFs fail to capture key information about the statistical covariance of environmental analyte concentrations. Specifically, it can generally be shown that surface water ion concentrations are not statistically independent, *i.e.*, the joint PDF is not simply related to the product of the individual ion PDFs:[29]

$$p_{x_1,x_2}(x_1, x_2) \neq p_{x_1}(x_1)p_{x_2}(x_2) \tag{1}$$

for any two concentrations for the ions studied here, and by extension for the entire suite:

$$p_{x_1\cdots x_n}(x_1\cdots x_n) \neq \prod_i^n p_{x_i}(x_i) \tag{2}$$

Open Access Article. Published on 03 May 2016. Downloaded on 11/18/2025 5:11:36 PM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

To accurately capture the covarying properties of the target ions in natural water samples, it is therefore necessary to refer instead to the distribution described by the n-dimensional joint PDF $p_{x_1 \ldots x_n}(x_1 \ldots x_n)$.

USGS data were used to create a discrete estimate of this joint PDF by identifying instances where surface water ions had been measured simultaneously (identical sample date, time, and site). The total set of approximately 200 000 available data points yielded 3218 instances when simultaneous measurements of the full set of 8 ions were made. The joint PDF was represented as an 8-dimensional matrix (each dimension representing the concentration of a single analyte) such that each entry in the 8-D matrix specified a concentration range (dictated by the binning width) of each of the 8 analytes. Concentration ranges for each analyte were individually divided into 10 bins equally spaced in $\log_{10}$[M] units. The 3218 points identified above were indexed (the 8-D index corresponding to the bin number on each axis), the total number of samples indexed to each location counted, and the final counts divided by the total (3218) to produce an 8-D surface which encloses a hypervolume of 1.

2.4 Statistically representative training samples

Based on this statistical model, a representative set of training samples was generated for ANN training. Seventy-five environmentally representative sample compositions were selected randomly from the 8-D joint PDF using a Monte Carlo methodology for discrete random variables.[30] The concentration for each constituent was selected from a uniform random distribution across the selected bins; however ammonium and nitrate were specified at 'low' (3 μM) and 'high' (100 μM) levels in a subset of samples to support the particular study of quantification of these analytes.

Synthetic samples were created from sixteen stock aqueous solutions (NaCl, Na_2SO_4, Na_2CO_3, KCl, KNO_3, K_2CO_3, $CaCl_2$, $Ca(OH)_2$, $MgCl_2$, $Mg(NO_3)_2$, $MgSO_4$, $MgCO_3$, NH_4Cl, HCl, HNO_3, and H_2SO_4). All solutions were 100 mM concentration, with the exception of $Ca(OH)_2$ and $MgCO_3$ which were 20 mM and 1.2 mM, respectively, due to their low solubility. Except for the HCl standard which was diluted from a 0.1 N aqueous standard, all standards were created using reagent grade salts, dried overnight at 55 °C if anhydrous or purchased new for hydrated salts, and weighed using an Ohaus precision standard TS4KD balance. Salts were dissolved in Millipore Milli-Q water (18.2 MΩ cm^{-1}) and diluted to the appropriate volume (typically 2 L) in a class A volumetric flask. Glass and plasticware used in this process were first acid washed for at least 24 hours in 10% HNO_3 and rinsed 7–10 times in Milli-Q water. Volumes of liquid stocks required to match the ion concentrations for the 75 samples were calculated. Specified volumes were added to Milli-Q water, diluted to 2 L in a class A volumetric flask, well mixed, and then transferred to 2 L LDPE bottles (acid cleaned and rinsed using the method specified above, after which they were capped and stored until use). Propagated errors due to weighing (±0.01 g) and dilution (±0.5 mL flask accuracy) bound concentration errors in the final samples at ≤±0.8 μM (with the highest relative errors expected for salts with low molecular weights or ions at low concentrations). pH of the resulting samples ranged from 7.1–8.6 when equilibrated with atmospheric CO_2. The electrical conductivity of samples ranged from 29 to 1644 μsiemen cm^{-1}.

2.5 Single-salt training samples

For comparison to the proposed statistically representative training samples outlined above, five sets of single-salt calibration standards were used to characterize the response of each ISE to each ion in the synthetic samples. Concentrations from 0.1 μM to 100 mM were used to span ranges identified in natural waters, with three standards per decade (*e.g.*, 1.0, 2.5, and 5.0 μM). The salts used were KNO_3, Na_2SO_4, $Mg(NO_3)_2$, NH_4Cl, and $CaCl_2$, and the same procedure was followed for standard creation and storage as described above.

2.6 Electrode selection and characterization

To evaluate ISE selections, the response of each ISE was measured independently for each of the ions considered in this experiment using the five sets of single-salt standards described above. Calibration curves were created for each ISE relative to its primary ion (or secondary ion, in cases such as Mg^{2+} where no ISE was available for a given ion) to provide a baseline method for sensor signal interpretation and quantification of interfering analyte contribution to signals in complex solutions. The linear response region was identified by maximizing R^2 of the linear fit for a variable number of calibration points, which provided an objective measure of where the 'knee' started. Sensors maintain a measurable response above the baseline well into the ~μM range, making these sensors theoretically usable at environmental levels (see Table SI.2†).

2.7 Creation of training datasets using training samples

The sensor response to training samples was measured as previously described[31] and is briefly summarized here. Electrodes were pre-conditioned following the manufacturers' recommendations (typically 10 minutes to 1 hour) once at the start of each sampling session. Electrode outputs were individually amplified using custom differential amplifiers based on the LMC6001 ultra-low input current op-amp (input resistance > 1 tera-ohm); op-amp outputs were connected directly to the giga-ohm input impedance analog inputs of the National Instruments 6218 data acquisition board (16 bit ADC, 250k samples per second). Custom LabView software recorded ISE potentials at approximately 1 Hz, with steady state potential identified following most recent IUPAC recommendations,[32] *i.e.*, with 'steady state' determined when the absolute value of the time derivative of the emf remains below a specified limit for a specified duration of time. The relevant values used were 0.4 mV min^{-1} and 40 s, as determined to be optimal in a previous study by these authors.[31] Response time of electrodes was 1–5.5 min depending on analyte concentration.

Seven replicates of each sample were measured, with measurements being made in a pseudo-random order relative to constituent concentrations for environmentally representative samples or with increasing concentration for single-salt standards. After ISE data were logged, electrical conductivity

Open Access Article. Published on 03 May 2016. Downloaded on 11/18/2025 5:11:36 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

(EC) (Amber Science Model 604) and temperature of each sample/replicate were measured. These measurements were not taken simultaneously with the ISE measurements to avoid interference to ISE signals by currents induced in the water during EC measurements. Electrical conductivity measurements were corrected for temperature following the literature.[33] The EC meter and pH ISE were calibrated daily with commercial calibration standards (0.73–10 000 μsiemen cm^{-1} at 25 °C; pH 4, 7, 10). Calibration at the end of the sampling period (approximately two weeks) was not statistically different compared to the initial calibration, *i.e.*, the slope and intercept were within the confidence interval of the initial linear calibrations (drift was negligible).

2.8 ANN software calibration and optimization

Development and optimization of the novel extended artificial neural network (ANN) architecture used for this work is described in detail in previous publications.[19,34] Briefly, ion concentrations in training samples and corresponding hardware responses were used to train a wide range of potential ANNs from which the optimum was ultimately selected for each combination of training set and architecture constraints. Sensor outputs (11 ISE mV outputs, one EC value, and temperature) served as the input data, while the known ionic concentrations, the quantities of interest for this analysis, in each of the samples served as the targets. Absolute concentrations [mol L^{-1}] were used as targets rather than the corresponding $\log_{10}$ data (use of log-transformed data was investigated as a technique often useful when concentration ranges span several orders of magnitude) as this was shown to decrease error by ≥50% for this application. Chemical constraints based on charge balance and electrical conductivity were implemented following previous work.[19]

To test the hypothesis that the statistical representativeness of training samples affects the achievable quality of software optimization, ANNs were trained with several types of training datasets measured simultaneously by the full sensor array: (1) single-salt standards spanning the relevant concentration space (termed 'SS'), (2) statistically representative synthetic environmental samples (termed 'SR'), and (3) a dataset comprised of both (1) and (2) (termed 'SS + SR'), with the degree of statistical representativeness increasing from SS to SS + SR to SR. To test the effect of integrating chemical understanding into the ANN architecture, ANNs were trained with this capability alternatively included and excluded, the latter being standard practice for ANN techniques and therefore serving as a baseline reference. In each case, the optimal ANN was selected by minimization of NRMSE on the 8 ion channels, with a 10 : 1 weighting on the nitrate and ammonium channels used to co-optimize results for ion concentrations ranging across several orders of magnitude.

3 Results

3.1 Statistical character of New England waters

Water quality data tend to be managed by individual states, making it convenient to compare one-dimensional PDFs for several key ions by state (MA, CT, VT, NH, ME) – see Fig. 1. Note that binning was conducted by dividing the observed ranges into 10 equal-width bins in $\log_{10}$[M] units to create these discretized approximations of the true underlying PDFs. The accuracy and resolution of such estimated distributions are dependent on the sample size; in this study, the number of available samples was relatively large ($300 < n < 3800$ for each state/ion combination), allowing adequate resolution of ion PDFs at the 10-bin level.

Examination of calcium concentration, as well as the Ca^{2+}/Na^{+} ratio, shows that waters in these five states range from hard to soft: Vermont waters had significantly higher calcium concentrations and relatively lower sodium concentrations than other states (a fact also reflected in pH), while waters from Connecticut are more frequently soft. Waters from New Hampshire show a strongly bimodal distribution for calcium content. For all states considered, mean values of nitrate and ammonium concentrations are approximately 10 μM, and distributions show the majority of samples within the 0.1–10 μM range, corroborating prior literature stating that concentrations of tens of μM are of interest for natural systems. The nitrate and ammonium distributions also vary by state, presumably reflecting both ecological and land use conditions. Electrical conductivity is constrained to a range from 10–3000 μsiemen cm^{-1}. Overall single-ion PDFs tend to be skewed and/

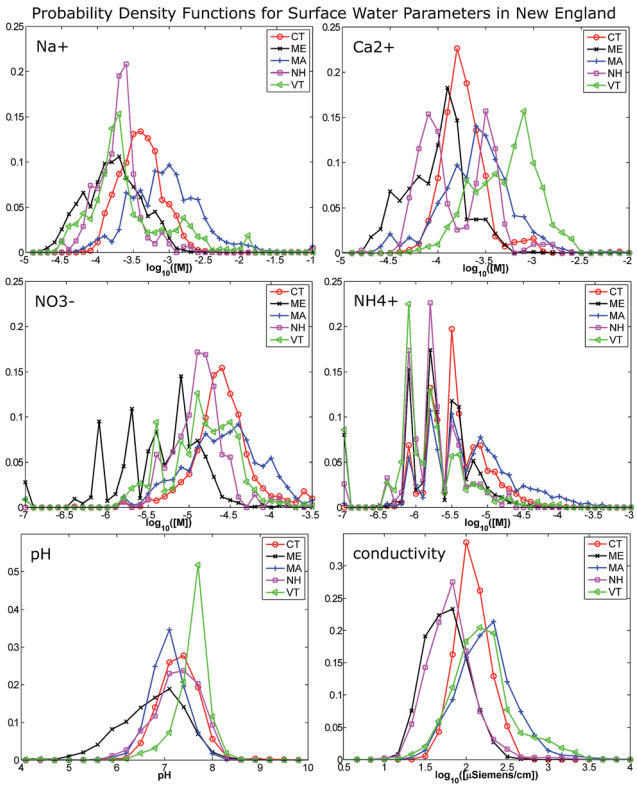

Fig. 1 One-dimensional probability density functions for representative environmental ions and parameters, created using archived USGS data[28] for the five states listed. Density values are plotted at bin midpoints. Substantial variability across the geography of New England motivates the necessity of creating a training set which encompasses data from all states.

Open Access Article. Published on 03 May 2016. Downloaded on 11/18/2025 5:11:36 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

or multi-modal; ranges and averages calculated from the data for these and other key ionic constituents (combined for all five states) are given in Table 1 where, for reference, published and measured response ranges for commercial solid state ISEs are also provided.

3.2 Statistical relationship of ion concentrations

The extent to which analyte concentrations covary was investigated through an analysis of the shape/distribution of the 8-D joint PDF, *i.e.*, the 8-dimensional version of a histogram (such as those given in Fig. 1) wherein the number of counts falling into any particular bin represents the proportional likelihood of sample concentrations falling within the corresponding concentration ranges. In the 8-D joint PDF, only 152 (of 10^8) bins had counts ≥ 5 and a full 80% of sample density was represented by only 401 bins (~0.0004% of the total 8-D hypervolume), whereas 80% of the density of an 8-D multivariate normal distribution would be concentrated in approximately 8% of the hypervolume. The data thus indicate that there is a high degree of covariance in the environmental joint PDF. To partially visualize this phenomenon, a subset of 2-D joint PDFs are presented in Fig. 2.

The top two panels of Fig. 2 show combined effects of weathering, runoff (*e.g.*, including road salts), and groundwater/precipitation mixing, resulting in an overall 'more is more' trend in the data, *i.e.*, a distribution that is strongly skewed along the positive slope (non-Gaussian in nature). In such cases, the conditional PDF, here $p_{K^+|Na^+}(K^+|Na^+)$, is a strong function of sodium concentration. The mid-level panels show the joint distributions of nitrate with chloride and calcium, while the lower panels show joint distributions of ammonium with sodium and nitrate, to demonstrate that these relationships hold even for nutrients which would not be expected to be jointly produced *via* weathering or necessarily be present in a predictable ratio with the other water constituents.

Overall it is clear that, in addition to the single-ion PDFs being skewed or multi-modal, the multidimensional joint PDFs are neither uniform nor well represented by a multivariate normal distribution. This high degree of correlation in environmental ion concentrations supports the hypothesis that constant background samples with varying additions of ions of interest are unlikely to statistically honor the actual ion distributions seen in target waters. It also supports our hypothesis that a signal processing method can take advantage of the additional information available in these statistical relationships (*e.g.*, artificial neural networks, which integrate such information during the training process). Finally, PDFs contain information that may be useful in the formulation of hypotheses regarding control of regional water chemistry.

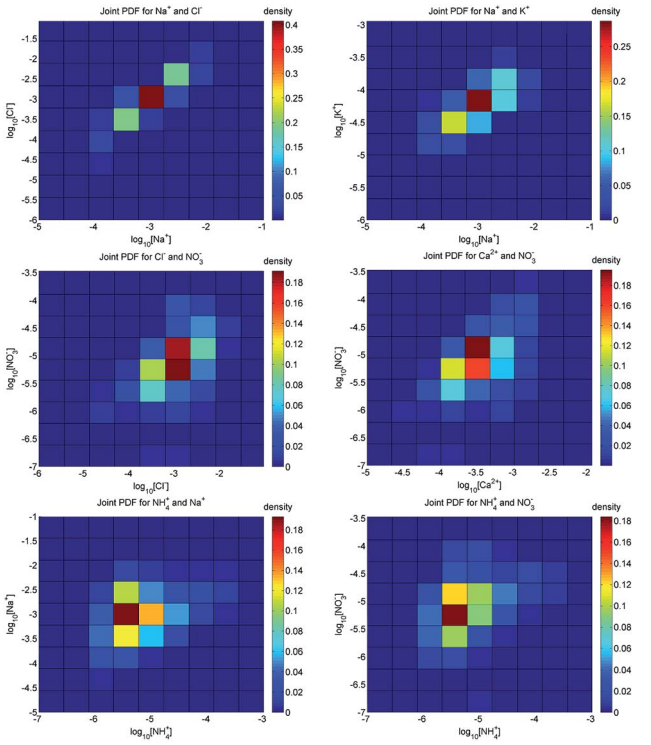

Fig. 2 Two-dimensional joint probability density functions for various ionic constituents of natural waters, created using archived USGS data for the five states listed. Density values are plotted at bin mid-points. Distributions demonstrate an overall tendency for ion concentrations to be positively correlated, however scatter around this line is typically up to two orders of magnitude in width producing substantively irregular distributions.

Table 1 Approximate concentration ranges for ions of interest in New England waters surveyed by the USGS over the past 50 years ($\log_{10}$[M]).[28] The commercial solid state ISE manufacturer-published 'range limit' and measured LOD are provided for comparison where available (from Tables SI.1 and SI.2)

Analyte	Min.	Mean	Max.	Range (manuf.)	LOD (meas.)
NO_3^-	−6.8	−5.1	−3.7	−5.3	−5.5
NH_4^+	−6.8	−5.0	−3.2	−5.7	<−6.0
Na^+	−4.8	−2.7	−1.2	−5.7	−6.3
Ca^{2+}	−4.8	−3.5	−2.2	−6.3	−5.9
K^+	−5.8	−4.3	−3.2	−5.0	<−6.6
Cl^-	−5.7	−2.8	−1.3	−4.5	−5.5
Mg^{2+}	−5.7	−3.9	−1.3	N/A	N/A
SO_4^{2-}	−5.3	−3.8	−1.3	N/A	N/A

3.3 Integration of statistical knowledge to improve *in situ* ISE use

Benefits of integration of this statistical understanding of the target waters into the design of *in situ* instrumentation were evaluated by training two types of ANNs (standard and extended with chemical knowledge) each with three different sets of training data having varying degrees of statistical representativeness of natural surface waters (SS, SS + SR, SR), as described earlier. A standard ANN architecture trained with single-salt standards bracketing concentrations of interest for each analyte is taken as the base case for the measurement of relative improvement as the proposed strategies are progressively added. Note that ANN architectures are generally expected to improve in predictive capability as the number of training

Open Access Article. Published on 03 May 2016. Downloaded on 11/18/2025 5:11:36 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

samples is increased, and therefore, the SS + SR and SR cases represent a measure of the advantages of 'quantity' *versus* 'quality'. While the instrumentation quantifies all 8 target ions[19,34] (the two nutrient species of interest and 6 other ions identified as important due to direct cross-reactivity with NH_4^+ or NO_3^- ISEs), the discussion presented here will focus primarily on results for the nitrogen species as these are of particular environmental interest.

Fig. 3 demonstrates the extent to which interference is experienced by the nitrate and ammonium ISEs with no signal post-processing. It is clear that the use of a linear calibration alone will fail to usefully determine actual ionic distribution in the samples; errors in both nitrogen species as estimated using these curves (representing the naive use of these ISEs as stand-alone sensors as one would a pH sensor) further illustrate the challenge in adapting these sensors for *in situ* use (illustrated in more detail below). A salient, though not unexpected, aspect of the results is the large offset bias for ammonium or nitrate that is introduced when using a simple linear calibration for samples with relatively high concentrations of interfering ions.

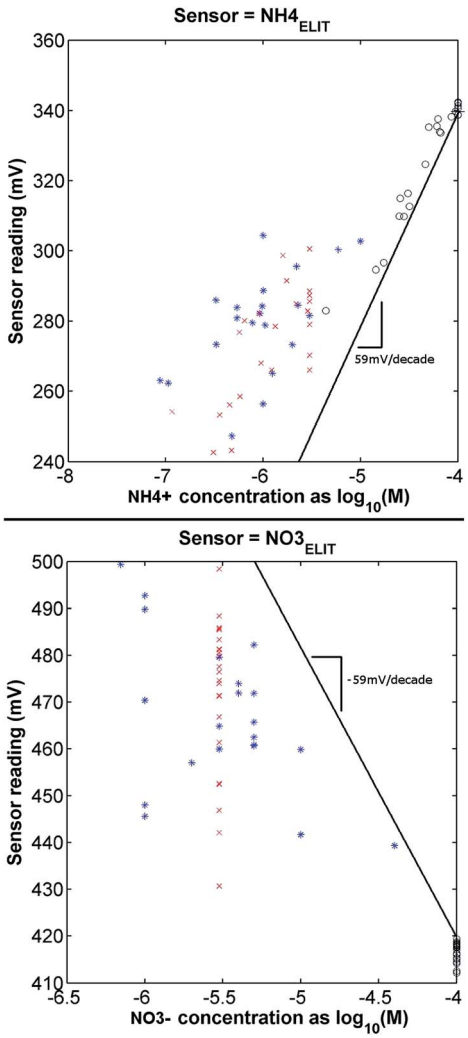

Fig. 3 Mean response of N-specific ISEs as a function of primary analyte concentration for environmentally statistically representative synthetic training samples. The theoretical Nernstian slope (approximately 59 mV per decade) is shown for visual reference. The vertical spread around what would be expected to be a linear calibration can be interpreted as interference leading to approximately a factor of 10 error in prediction of these species at typical environmental levels (≤10 μM).

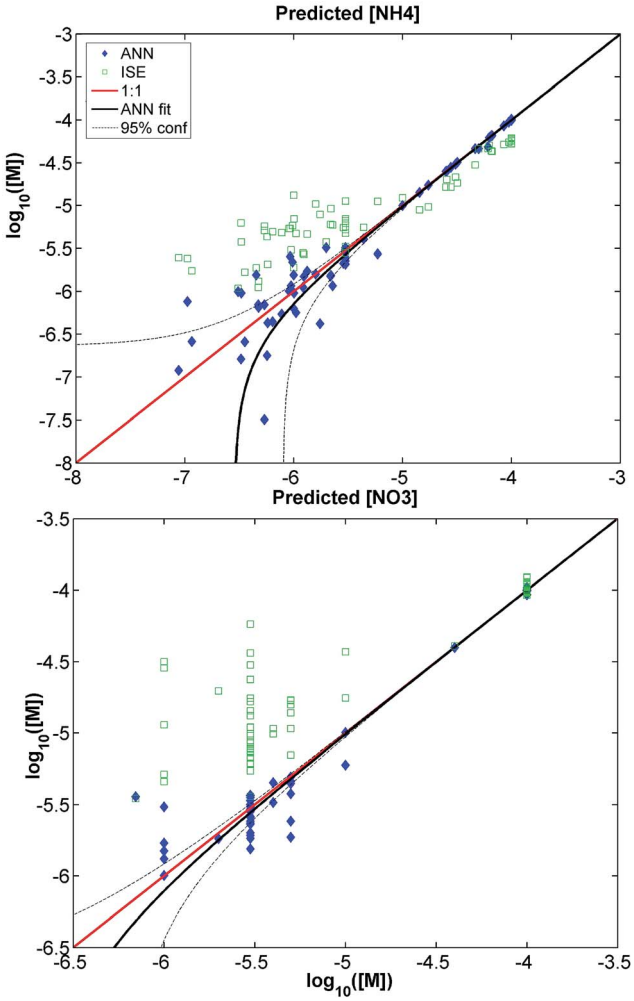

Fig. 4 Scatter plots of NH_4^+ and NO_3^- ion concentrations estimated using the optimal ANN (extended ANN trained with set SR) as a function of true concentration (blue filled diamonds) as compared to results using ISE linear calibrations (green open squares) where effects of interfering analytes are highly visible. One-to-one line shown in red; regression of ANN estimates against targets (concentration data) and 95% confidence interval on the linear fit shown in black.

3.4 N species estimation: optimal signal processor

Estimates of nitrate and ammonium concentrations from the optimized expanded-ANN architecture (using the optimal extended ANN configuration identified in ref. 19 and trained with the most representative samples as described above) were compared against the best results achievable using a standard ANN configuration and two other training sets with decreasing statistical representativeness of natural waters.

Table 2 shows a comparison of the estimation errors (NRMSE) for the two types of ANNs and three training sets,

Open Access Article. Published on 03 May 2016. Downloaded on 11/18/2025 5:11:36 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Table 2 Ion concentration estimation errors as normalized root mean square error (NRMSE) for standard and extended ANNs as trained with (I) only single-salt standards (SS), (II) both single-salt standards and statistically representative samples (SS + SR) and (III) only statistically representative samples (SR). NRMSE is given for ammonium ion (NH_4^+), nitrate ion (NO_3^-), and the summation of NRMSE for the 8 target ions (ALL) (note: data for the Extended/SR case shown in Fig. 4). Factor of improvement is shown disaggregated by improvement due to statistical representativeness of data (progression from SS to SR) and integration of chemical knowledge into the ANN algorithm (standard to extended)

	Training set	NRMSE		
		NH_4^+	NO_3^-	ALL
Standard ANN	SS	1.742	1.128	10.317
	SS + SR	0.139	0.098	1.911
	SR	0.271	0.097	2.02
Extended ANN	SS	1.514	1.257	9.915
	SS + SR	0.201	0.082	1.63
	SR	0.081	0.037	1.024
		Factor of improvement		
Standard ANN, SR *vs.* SS		6.43	11.63	5.11
Extended *vs.* standard ANN, SR		3.35	2.62	1.97

disaggregated as NRMSE for ammonium, nitrate, and the sum for all 8 studied ions. The comparison of results with training sets SS, SS + SR, and SR demonstrates the contribution of improving the statistical representativeness of the training data: total error is decreased by a factor of 5, while error on nitrate/ammonium channels decreases by a factor of 6–12. Notably, comparing the SR training results against SS + SR results provides a measure of the value of quantity *vs.* quality for ANN training; in the extended ANN case, the inclusion of SS training samples actually decreases predictive capability. The comparison of extended and standard ANN architectures trained with the SR set provides a measure of the improvement achieved due to the integration of chemical knowledge into the ANN architecture: total error is decreased by a factor of 2, while nitrate/ammonium error decreases by a factor of 2.6–3.4. It is furthermore clear that synergies are achieved by implementing both improvements simultaneously; the extended ANN algorithm provides an increasing improvement relative to standard ANNs as the training set increases in representativeness. For reference, the NRMSE values for the optimal ANN estimates correspond to mean relative errors of 10–20% for nitrate and ammonium even at concentrations as low as or lower than 10 μM, while the standard ANN trained with the SS set has MRE > 450% and the standard ANN trained with SR has a MRE of 30–55% for nitrate and ammonium.

Scatter plots in Fig. 4 show the results achieved using the optimal ANN. ISE log-linear calibration curves were also used to produce concentration estimates (expected to be highly erroneous due to the presence of interfering analytes) which are overlain for visual reference of the level of experienced interference. Note that vertical scales are different for each subplot and are dictated by natural water concentrations (discussion above). Significantly, in both cases the optimal ANN produces concentration estimates (blue, filled) that generally fall along the 1 : 1 lines (red), having successfully removed the bias that is encountered when the ISEs are used in multicomponent mixtures (demonstrated by the displacement of the bulk of open, green points from the 1 : 1 line).

Table 3 Parameterization of the linear regression of optimal ANN (extended architecture, SR training set) derived concentrations against target concentrations for nitrate and ammonium; in both cases, the slope and intercept are statistically indistinguishable from 1 and 0, respectively

	NH_4^+	NO_3^-
Slope	0.996 ± 0.013	1.000 ± 0.007
Intercept	$(-2.84 \pm 5.14) \times 10^{-7}$	$(-2.15 \pm 4.34) \times 10^{-7}$
R^2	0.997	0.999
RMSE [M]	1.77×10^{-6}	1.38×10^{-6}

Table 3 shows the parameters for linear fits of the optimal ANN-derived (extended architecture, SR training set) concentrations against the target concentrations. The 95% confidence interval on the slope contains 1 (perfect agreement) for both nitrate and ammonium, and intercepts are not statistically significantly different from zero. These facts identify the extended ANN as an unbiased estimator (*i.e.*, accurate), and therefore, both the R^2 and RMSE values provide information about the magnitude of scatter around the targets (precision). It is significant to note that concentration estimates of both nitrate and ammonium were unbiased even for points below the published detection limits of their respective electrodes.

4 Conclusions

A statistical analysis of natural water chemistry data demonstrates that ion distributions are strongly correlated yet highly irregular in the New England region and confirms that detection limits of ≤10 μM are needed to appropriately characterize nitrate and ammonium levels in a large fraction of New England surface waters. This appears to be feasible using ISE arrays with ANN post-processing, however it is clear that neither sets of single-salt standards bracketing expected concentration ranges nor collection of environmental samples are likely to provide a training set that is adequately statistically representative of target waters. In such cases, the proposed methodology (creation of synthetic training samples based on a joint statistical model of ion concentrations in environmental surface waters) provides a valuable tool for expediting development and training/calibration of sensor instrumentation. From a geochemical perspective, the statistical relationships among ions may also suggest hypotheses regarding underlying controls of ionic concentrations in different watersheds throughout New England as well as provide information which can be utilized by the signal processing module to lower detection limits for ions of interest.

Open Access Article. Published on 03 May 2016. Downloaded on 11/18/2025 5:11:36 PM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Nitrate and ammonium concentrations are quantified in unprocessed samples at environmental levels and in environmentally representative multi-analyte solutions, providing unbiased concentration estimates down to <10 μM, with improvements coming both from increasing statistical representativeness of the training set and integrating chemical constraints into the ANN architecture. Mean relative errors of ≤20% of absolute concentration are achieved for most samples (with maximum system errors of approximately 50% at the micromolar concentrations for NH_4^+). This lowers detection limits by more than an order of magnitude relative to those achieved by current commercial state-of-the-art ISE-based instruments for nitrate and ammonium (>100 μM) and removes systematic bias at these low concentrations. These are two major steps in the direction of real-time *in situ* quantification of these environmentally important analytes; these capabilities are critical for eutrophication studies, particularly in watersheds that are strong contributors to estuaries where nitrogen is typically suspected to be a major contributor.

The hardware used in this study, including the ISE array, signal processing circuitry, and computational capability implementing the ANN, indicates that it is possible to implement such a system in a package which is reasonably light and compact (<15 lbs) and has low power requirements (≪1 W for sensors and signal conditioning, with overall power consumption dominated by the single board computer (1–5 W)). Each measurement requires ∼5 min, making such a system an excellent candidate for field use. The short measurement time and capability for large numbers of measurements in a single deployment would further enable adaptive sampling campaigns, which may also aid in the identification of broad non-point sources of nitrogen nutrients to surface waters through enabling real-time 'tracking' of nutrient signals. Even when simply used as a guide for placement of limited grab samples, the capability for real time sampling can improve sample density, decrease uncertainty about the true characteristics of a given ecosystem, and provide key data necessary to inform environmentally conservative and fiscally responsible management decisions. This study, therefore, presents an important step in the direction of development of *in situ* instrumentation for the broad support of scientific field studies.

A number of practical issues remain to be addressed to facilitate the widespread use of the proposed hardware/software architecture for scientific field work. Among these are an appropriate daily calibration routine to counter the effects of sensor drift (*e.g.*, with 3–5 multi-analyte standards), a formal analysis of the tradeoff between the number of sensors and achievable accuracy (noting that error increased with removal of any sensors included in this study), and a method for generalizing application to other regional waters. The last is of particular interest as not all regions have historically been extensively characterized, and therefore a system capable of adapting the calibration for use across a wide range of environments would be highly desirable.

Enhancement of the results presented here could further be achieved through development of ISEs for sulfate, magnesium, and/or carbonate at appropriate levels (none of which currently exist), as the nature of the ANN architecture takes advantage of information available in all input signals to improve the accuracy of all output signals. Expansion of the sensor set to include DO and Fe would enable characterization of a wider range of surface waters. Additional research directions include further optimization (and potential expansion) of the training set and examination of possible effects of the background matrix of real waters; in particular, it is suggested that the effects of humic materials, inevitably present in natural waters, be explored. While humic acids are not necessarily expected to affect the response of ISEs themselves, they will contribute to the overall charge balance of natural waters and their effects will therefore need to be incorporated into the ANN's chemical model. To overcome effects of humic acids, addition of optical channels may be warranted and could potentially provide additional information for other analytes (*e.g.*, nitrate). Finally, due to the strong interest in the eventual consequences of nitrogen contributions to estuarine and coastal systems, exploration of the effectiveness of the proposed methodology for more saline waters is recommended.

References

1 I. Valiela and J. Bowen, *Environ. Pollut.*, 2002, **118**, 239–248.
2 G. McIsaac, M. David, G. Gertner and D. Goolsby, *Nature*, 2001, **414**, 166–167.
3 R. Howarth and R. Marino, *Limnol. Oceanogr.*, 2006, **51**, 364–376.
4 K. Johnson and L. Coletti, *Deep Sea Res., Part I*, 2002, **49**, 1291–1305.
5 YSI, *YSI 6820 V2 Compact Sonde for Field Sampling of Dissolved Oxygen and More*, 2013, http://www.ysi.com/productsdetail.php?6820-V2-4.
6 M. Cutrofello and J. Durant, *Chemosphere*, 2007, **68**, 1365–1376.
7 M. Otto and J. Thomas, *Anal. Chem.*, 1985, **57**, 2647–2651.
8 C. Di Natale, F. Davide, J. Brunink, A. D'Amico, Y. Vlasov, A. Legin and A. Rudnitskaya, *Sens. Actuators, B*, 1996, **34**, 539–542.
9 J. Mortensen, A. Legin, A. Ipatov, A. Rudnitskaya, Y. Vlasov and K. Hjuler, *Anal. Chim. Acta*, 2000, **403**, 273–277.
10 A. Rudnitskaya, A. Ehlert, A. Legin, Y. Vlasov and S. Buttgenbach, *Talanta*, 2001, **55**, 425–431.
11 C. Di Natale, A. Macagnano, F. Davide, A. D'Amico, A. Legin, Y. Vlasov, A. Rudnitskaya and B. Selezenev, *Sens. Actuators, B*, 1997, **44**, 423–428.
12 J. Gallardo, S. Alegret, R. Munoz, M. de Roman, L. Leija, P. Hernandez and M. del Valle, *Anal. Bioanal. Chem.*, 2003, **377**, 248–256.
13 M. Baret, D. Massart, P. Fabry, F. Conesa, C. Eichner and C. Menardo, *Talanta*, 2000, **51**, 863–877.
14 L. Duarte, C. Jutten and S. Moussaoui, *8th International Conference on Independent Component Analysis and Signal Separation*, 2009, pp. 662–669.
15 P. Dillingham, T. Radu, D. Diamond, A. Radu and C. McGraw, *Electroanalysis*, 2012, **24**, 316–324.
16 M. Gutierrez, S. Alegret, R. Caceres, J. Casadesus, O. Marfa and M. del Valle, *Comput. Electron. Agr.*, 2007, **57**, 12–22.

 This journal is © The Royal Society of Chemistry 2016

Open Access Article. Published on 03 May 2016. Downloaded on 11/18/2025 5:11:36 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

17 M. Gutierrez, S. Alegret, R. Caceres, J. Casadesus, O. Marfa and M. del Valle, *J. Agric. Food Chem.*, 2008, **56**, 1810–1817.
18 M. Gutierrez, J. Gutierrez, S. Alegret, L. Leija, P. Hernandez, L. Favari, R. Munoz and M. del Valle, *Int. J. Environ. Anal. Chem.*, 2008, **88**, 103–117.
19 A. Mueller and H. Hemond, *Talanta*, 2013, **117**, 112–118.
20 M. Bos, A. Bos and W. van der Linden, *Analyst*, 1993, **118**, 323–328.
21 P. Daponte and D. Grimaldi, *Measurement*, 1998, **23**, 93–115.
22 M. Bos, A. Bos and W. van der Linden, *Anal. Chim. Acta*, 1990, **233**, 31–39.
23 M. Cortina, A. Gutes, S. Alegret and M. del Valle, *Talanta*, 2005, **66**, 1197–1206.
24 E. Richards, C. Bessant and S. Saini, *Chemom. Intell. Lab. Syst.*, 2002, **61**, 35–49.
25 J. Gallardo, S. Alegret, M. de Roman, R. Munoz, P. Hernandez, L. Leija and M. del Valle, *Anal. Lett.*, 2003, **36**, 2893–2908.
26 F. Despagne and D. Massart, *Analyst*, 1998, **123**, 157–178.
27 R. Srivastav, K. Sudheer and I. Chaubey, *Water Resour. Res.*, 2007, **43**, 1–12.
28 United States Geological Survey, *USGS Water Quality Samples for USA: Sample Data*, 2010, http://nwis.waterdata.usgs.gov/nwis/qwdata.
29 W. Mendenhall, R. Beaver and B. Beaver, *Introduction to Probability and Statistics*, Duxbury Press, Pacific Grove, CA, 14th edn, 2012.
30 C. Robert and G. Casella, *MonteCarlo Statistical Methods*, Springer Science, New York, NY, 2nd edn, 2004.
31 A. Mueller and H. Hemond, *Anal. Chim. Acta*, 2011, **590**, 71–78.
32 R. Buck and E. Lindner, *Pure Appl. Chem.*, 1994, **66**, 2527–2536.
33 M. Hayashi, *Environ. Monit. Assess.*, 2004, **96**, 119–128.
34 A. Mueller, PhD thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, 2012.