Instrument bias correction with machine learning algorithms: Application to field-portable mass spectrometry

B. Loose, R.T. Short, S. Toler

In-situ sensors for environmental chemistry promise more thorough observations, which are necessary for high-confidence predictions in earth systems science. However, these can be a challenge to interpret, because the sensors are strongly influenced by temperature, humidity, pressure, or other secondary environmental conditions that are not of direct interest. We present a comparison of two statistical learning methods - a Generalized Additive Model, and a Long-Short Term Memory (LSTM) Neural Network model for bias correction of in-situ sensor data. We discuss their performance and tradeoffs, when the two bias correction methods are applied to data from submersible and shipboard mass spectrometers. Both instruments measure the most abundant gases dissolved in water, and can be used to reconstruct biochemical metabolisms, including those that regulate atmospheric carbon dioxide. Both models demonstrate a high degree of skill at correcting for instrument bias using correlated environmental measurements; the difference in their respective performance is less than 1% in terms of root mean squared error. Overall the LSTM bias correction produced an error of 5% for O_2 and 8.5% for CO_2, when compared against independent membrane DO and laser spectrometer instruments. This represents a predictive accuracy of 92-95% for both gases. It is apparent that the most important factor in a skillful bias correction is the measurement of the secondary environmental conditions that are likely to correlate with the instrument bias. These statistical learning methods are extremely flexible and permit the inclusion of nearly an infinite number of correlates in finding the best bias correction solution.

1.0 Introduction

The uncalibrated signal (s) produced by an environmental sensor contains the superposition of multiple influences. These include the instrument response to an environmental property of interest: $y(\vec{x}, t)$, but it also includes some instrument responses (β) that are not of interest, as well as some uncorrelated or random error (ϵ). The undesirable influences in β can be represented if the environmental influences or correlates, X, are separately measured. An example of $\beta(X)$ would be changes in the internal resistance of a circuit board as the room temperature varies. We refer to $\beta(X)$ as instrumental bias, and their influence on s can be treated as additive

$$s = y(\vec{x}, t) + \beta(X) + \epsilon \quad (1)$$

and therefore, separable from $y(\vec{x}, t)$, the desired environmental response.

Experimental chemistry has been slow to consider bias and systematic error, in part because the end goal of many studies was the demonstration of a corollary relationship, rather than a process model (Newman, 1993). However, when the same relationships are used in a predictive capacity, the uncorrected bias can lead to erroneous results. Recently, bias has been

given more explicit treatment, through applications such as air quality for human health (Delle Monache et al., 2006) and charge state in electric vehicles (Sun et al., 2016). These and other application demand accurate forecasts, thereby renewing focus on elimation of bias from the process model.

Within the geosciences, the problem of chronic under-sampling in diffusive environments, such as air and water, (Pimentel, 1975) has created a strong incentive to take instruments out of the lab to increase sample density and better characterize the tracer field. If samples are analyzed in a discrete fashion, instrumental drift that leads to bias, can be accounted for with pre/post calibration to constrain the instrument drift. This was the approach adopted by e.g. Guegen and Tortell (2008) to measure dimethyl sulfide (DMS) and carbon dioxide – two climatically important gases – during a shipboard expedition in the Southern Ocean. However, the continuous sampling that takes place with in-situ or underway chemical sensors, requires a slightly different approach to account for instrument drift as a source of bias. One clever solution has been to switch to reference compound(s) at regular intervals as part of the measurement protocol. This has the effect of chopping up the time series and introducing data gaps, but these gaps are often small (minutes) in comparison to the averaging interval (tens-of-minutes to hours) that is utilized for final data presentation. Takahashi et al., (2002, 2009) have used the approach of reference compounds at intervals to create very precise coverages of ocean surface carbon dioxide concentration for several decades. Cassar et al., (2009) showed that mass spectrometer drift, while measuring oxygen and argon, could be characterized by switching regularly to measure atmospheric air. Saltzman et al., (2009) describe a detailed method for measurement of continuous DMS measurements using a Chemical Ionization Mass Spectrometer. Their approach, which uses DMS isotope dilution, also uses switching at intervals to characterize several bias corrections and account for internal sources of DMS, as well as sensitivity of the instrument to changes in seawater temperature and other environmental factors. These biases are reported at less than 1% of the overall DMS signal.

The approach of regular switching to a reference compound is a proven means to correct for drift in continuous instruments. However, the instrumental conditions that we confront in this study differ in two significant ways from the previously-described continuous measurement methods. The first difference has to do with the magnitude of the bias, compared to the signal of interest. Previous underway studies have confronted bias corrections of a few to 10% of the overall instrumental signal, while the instrumental bias that we face can vary by 100% or more. The magnitude of this bias renders the true environmental signal unrecognizable, until the correction has been applied. The second major difference is that previous studies have identified the most likely sources of bias, but they have not quantified those sources to implement the bias correction. When the instrumental bias masks the true environmental signal, the bias must be treated as a continuously varying function and therefore a simple linear correction to baseline drift is not adequate. This bias correction problem lends itself to time series and multivariate regression techniques, including partial-least squares, ridge regression, generalized linear and generalized additive models (Hastie et al., 2001).

Multivariate time series predictions have undergone a period of rapid development and availability thanks to the popularity of another member of the statistical learning family - neural networks, which have proven facile at e.g. image and speech recognition. Neural networks are

also suited for time series applications including forecasting or prediction (Brownlee, 2019a). Specifically, the Long Short-Term Memory (LSTM) algorithm combines the learning power of neural networks with a capacity to down-weight or "forget" information that does not prove relevant, leading to the overall stability of the network optimization (Hochreiter and Schmidhuber, 1997).

In this application, we apply and compare a Generalized Additive Model (GAM) and a LSTM Neural Network model to observe their performance in baseline correction to mass spectrometer data. A schematic depiction of the bias correction workflow can be observed in Figure 1. Both the GAM and LSTM models use the statistical learning approach to optimize their calibration and weight coefficients. However, there is a fundamental difference in approach and user control. The LSTM weights and tradeoffs are largely abstracted from the user; one has to trust the algorithm without being able to interrogate the details of the solution. The consolation is the tremendous skill that the LSTM models exhibit in preserving the information that is necessary to discriminate or predict, while avoiding the spurious oscillations that can characterize simpler, stiffer models. Unlike the LSTM, the GAM represents a linear combination of regression models (Wood, 2017) between each environmental correlate (X_i) and the instrument signal (s). This allows the user to observe and evaluate the *partial dependence* of the GAM solution on each X_i and to alter the functional form (e.g. linear, polynomial, cubic spline) that is fit between s and each X_i. The effect is to give the user greater control over the functional form and the partial influence of each correlate on the total solution.

The signals of interest to this study are measurements of gases dissolved in water and seawater using field-portable quadrupole mass spectrometers (QMS). We present examples of the GAM and LSTM applied to data from a Submersible Wet Inlet Mass Spectrometer (SWIMS) that was used to measure dissolved oxygen in the top 200 m of the Sargasso Sea and Gulf Stream, in the subtropical Atlantic Ocean. We present a second example of signals collected with a similar mass spectrometer aboard a ship that was used to measure dissolved carbon dioxide at the ocean surface, within the sea ice-covered Ross Sea, Antarctica.

Throughout this text, we make references to the Python modules that were used to implement the individual solutions. The implementation of the GAM Backfit Algorithm, as well as example scripts for applying these methods to SWIMS data can be found in the Supplemental and in the Acknowledgements.

2.0 Methods

The bias correction models were each applied to ocean measurements of gases dissolved in seawater. These measurements were made using a Quadrupole Mass Spectrometer (QMS). The QMS is an ideal tool for ocean measurements, because it is compact, it can scan over a large range of atomic masses. In this study, we refer to the mass-to-charge ratio (m/z), where m represents the atomic mass of the molecule of interest and z represents the positive charge state. For example, water vapor is measured in the QMS at m/z = 18, and molecular oxygen (O_2) is measured at m/z = 32. In this study z = 1 in every instance. The QMS can be connected to a variety of gas inlet configurations. Further detail on the principles of quadrupole mass spectrometry can be found in (Dawson and Herzog, 1995), but they are not needed to follow the methods presented here.

2.1 Ocean data used to evaluate the bias correction models

SWIMS tow: The first ocean data set was collected in July, 2017 along a dynamic section of the subtropical Atlantic between 35° and 40° N latitude (Figure 2). The QMS was incorporated into a Submersible Wet Inlet Mass Spectrometer (SWIMS), which is capable of in-situ gas analysis to a water depth of 2000 m; in this application we towed the mass-spectrometer through water depths from 0 to 150 m aboard a Triaxus tow vehicle, corresponding to a region where sunlight penetrates the surface ocean. The SWIMS position can be visualized by the gray saw-tooth pattern in panel b of Figure 3. Calibration of the SWIMS instrument is described below in Section 2.2.

This ocean section began in the North Atlantic subtropical gyre a circulation feature that is known to be highly depleted of nutrients with low biomass (Jenkins, 1982). In summer, surface waters in the Gulf Stream and Gyre can exceed 30 °C, and nearly 1% of temperature measurements in this section fell between 30° and 35° C (Figure 3). North of the Gulf Stream, waters cool and become significantly fresher, reflecting river inputs and the influence of the southward-flowing Labrador Current (Chapman and Beardsley, 1989). We chose to test the bias correction models in this region, because the environment is highly changeable on a small horizontal and vertical scale, so the SWIMS is subjected to a wide range of environmental conditions, including temperature, salinity, and dissolved organic matter – all of which can cause the dissolved gas burden of the seawater to vary.

The SWIMS was being used to measure oxygen, argon, carbon dioxide, nitrogen and methane in the surface ocean. Each of these dissolved gases has significance for biology and geochemistry of the ocean. Our in-situ calibration system included reference gases for each of these compounds, allowing the SWIMS to reproduce realistic concentration distributions for each analyte. Here we will restrict analysis of the bias correction to the SWIMS signal at m/z = 32, corresponding to dissolved oxygen. By developing the bias correction at m/z =32, we are able to take advantage of independent measures of dissolved oxygen using a membrane oxygen sensor, the Seabird model SBE 43, which allows for a detailed reference time series, throughout the vehicle tow. Ultimately, we use the root-mean square error between the SBE43 and the SWIMS to establish a truly independent measure of the bias correction algorithm.

Shipboard QMS: The bias correction models were also tested on data from a shipboard QMS that continuously sampled dissolved gases in the Ross Sea sector of the South Atlantic, south of 75 °S. These measurements were collected between May 16 and June 4, 2017. The partial pressure of carbon dioxide (pCO_2) was measured by connecting the QMS directly to a turbulent air-water equilibrator of the type described by Takahashi (1961). The same equilibrator was used to measure pCO_2 by infrared absorption spectroscopy (Takahashi et al., 2002, 2009), again providing an independent measurement to compare the bias correction against. The QMS was connected to the equilibrator with a 2 m x 50 μm (len x dia) capillary, which served to throttle the gas flow into the QMS and thereby maintain a vacuum below 10^{-5} torr.

Carbon dioxide was measured with the QMS by scanning at the atomic mass m/z = 44. The reconstruction of pCO_2 was carried out with daily 3-point linear calibration with reference gases of pCO_2 = 0%, 0.4% 0.1%. These signals can be seen in the expanded scale on the right

side of Figure 4**Error! Reference source not found.**. Unlike the SWIMS tows, these calibrations were not long enough in duration to record the bias while sampling from a stable gas concentration. Therefore, we apply the bias correction to a time series of CO_2 partial pressure (pCO_2), measured at atomic mass m/z = 44. Instead, the GAM and LSTM models were trained on relatively stable ion current signals measured during a four-day period between May 27 and June 1.

This late autumn period in the Southern Hemisphere was cold and windy with continual disaggregated ice formation in the surface ocean. The principal source of bias appeared from the thermal cycling in the room where the QMS and equilibrator were operating **Error! Reference source not found.**(Figure 4). The heating system in that room would cause the temperature in the room to increase and decrease by 2-3 degrees Celsius every 30 minutes. Additionally, the seawater intake was periodically clogged with ice crystals, causing the equilibrator flow rate to vary.

2.2 In-situ calibration of the SWIMS

The SWIMS passes seawater directly over a gas-permeable silicone membrane under conditions that approach a constant flow rate, while maintaining constant water temperature using a resistive heater and aluminum block (Short et al., 2001; Wenner et al., 2004). The wet membrane inlet is a simple and elegant design that allows for a submersible instrument, but it is subject to a number of confounding environmental influences that complicate interpretation of the SWIMS ion current. The most significant of these is a change in membrane permeability as it is compressed under the increasing water pressure (Bell et al., 2007). The permeability behavior is made more complex by hysteresis between the compression and decompression cycles (Futó and Degn, 1994; Lee et al., 2016). Over progressive cycles the silicon membrane can become tempered and eventually exhibit less compressibility (Futó and Degn, 1994; Lee et al., 2016), which indicates that any bias correction should include multiple compression-decompression cycles to capture the longer-term transients. To capture this and other sources of bias, we designed an in-situ calibration method that involves connecting the SWIMS to a 1L Tedlar bag that contains seawater, equilibrated with a reference gas mixture. The sample in the compressible Tedlar bag is subjected to the same pressure variations as the water column sample, but gas concentrations remain constant, because there is no gas headspace in the bag. Using a 3-way solenoid switching valve, the SWIMS can change states from sampling the environment to sampling the constant reference gas. Because the gas concentration is invariant, any trends in ion current that are observed must be due to instrumental bias. An example of this instrumental bias can be observed in Figure 5, which shows the environmental correlates measured during approximately 4.8 tow cycles while measuring from the in-situ calibration reference. These signal variations are what we seek to correct.

2.3 Calibration after bias removal

To discover the instrument response, it is necessary to remove $\beta(X)$ the instrumental bias and rearrange equation (1) as follows

$$y(\vec{x}, t) = f(s - \beta(X)) \quad (2)$$

Here forward, we drop the explicit reference to uncorrelated error (ϵ), which means that this error source is still a part of *s*. After bias correction it is still necessary to estimate the uncertainty on *y* that is caused by ϵ, but that topic is extensively covered by other studies, so it will not be addressed here.

Therefore, the steps to obtain *y* are to first model $\beta(X)$ so that it can be removed and then to calibrate to obtain the empirical dependency, $f(\ \ )$, between *y* and the bias-corrected signal. To make this procedure less abstract, we focus on measuring the oxygen concentration in seawater $y = [O_2]$ using the ion current measured at m/z=32. The raw ion current (*s*) in milliamps at m/z = 32 responds directly to the amount of O_2 dissolved in the water, but also to other environmental correlates, *X*. The values of *X* must be measured as a time series, coincident with the the instrument's deployment. Other properties that we might include in *X*, are for example the duty cycle of a heater or chiller, the atmospheric pressure, the temperature of a chemically reactive solute (e.g. pH-sensitive dye), or the electrical conductivity of a water solution. The environmental correlates used to model $\beta(X)$ in the SWIMS are shown in Figure 5, and the correlates used to model $\beta(X)$ in the shipboard QMS are shown in Figure 4**Error! Reference source not found.**.

After bias removal, s reflects only the environmental signal of interest and some component of random error; $s_{-\beta}$ denotes the ion current after bias removal, and this term is calibrated against the reference gas concentrations using a linear equation,

$$f(s_{-\beta}) = m(s_{-\beta}) + b \text{ or} \qquad (3)$$
$$y(\vec{x}, t) = m(s_{-\beta}) + s^0$$

Here, the terms *m* and s^0 are the slope and intercept, and these terms are estimated as described in Section 2.1. Practically, we estimate *m* as,

$$m = \frac{y - y^0}{s_{-\beta} - s^0_{-\beta}} \qquad (4)$$

At the limit of $y^0 = [O_2] = 0$, the ion current does not reach zero because of electronic noise, and the potential for "virtual leaks" as gas is desorbed from the walls of the QMS under vacuum. In other words, y^0 is always zero, but in practice, $s^0_{-\beta}$ in equation (3) reflects the non-zero ion current at undetectable gas concentrations leaving the following linear calibration,

$$y(\vec{x}, t) = m(s_{-\beta} - s^0_{-\beta}) \qquad (5)$$

The technique for determining m and $s^0_{-\beta}$ for the Shipboard QMS were determined by measuring m/z = 44 at $pCO_2 = 0$ in ultrapure N_2 gas, as described in Section 2.1. The SWIMS determination of m occurred during in-situ calibration, which took place ca. every other day, however we did not determine $s^0_{-\beta}$ so it became necessary to account for baseline drift in the SWIMS using an external reference. We implemented a reference to the equilibrium oxygen solubility, based on seawater temperature and salinity. We also used the SBE43 as a daily reference.

2.4 General approach of statistical learning

The bias corrections that we evaluate here belong to a family of statistics called supervised learning. These corrections compare correlating *inputs* with corresponding *outputs*

to develop a *predictor*, that can be applied to any set of inputs. To develop the prediction, a sufficiently large dataset is divided into subsets - often referred to as 'train' and 'test' subsets (Ahmed et al., 2010). Separating in this manner allows the learning algorithm to develop a fit using the 'train' data set and evaluate the quality of that fit by *predicting* the data in the 'test' data set. The Scikit Learn module library in Python has been designed around the test-train convention and allows the user to subset using a number of different methods (Pedregosa et al., 2011). Last, the 'test' data set is used to estimate general error between the bias corrector and the actual data (Hastie et al., 2001).

Statistical learning models are exceedingly flexible and conform to almost any feature at any scale within a timeseries. This can result in 'overfitting', a condition where the learning algorithm attempts to reproduce small scale noise or other shapes in the data that do not improve the prediction or bias correction. Overfitting results because of the imperfect separation between the bias and the random error. This imperfect separation between β and ϵ, called the bias-variance tradeoff (Wood, 2017) results in a degradation of the fit as greater degrees of freedom are introduced to the model. Statistical learning algorithms included penalty parameters that can be adjusted to iteratively reduce the degrees of freedom. When this is done iteratively, one can probe the range of model-data misfit and determine the point where improved fitting becomes overfitting and then choose penalties accordingly in a process call Regularization (Hastie et al., 2001). We describe the application of penalty regularization to the GAM in Section 2.5 and to the LSTM in Section 2.6.

2.5 Implementation of the Generalized Additive Model and Backfit Algorithm.

A generalized additive model (GAM) achieves smooth fitting by using the sum of fitting functions that individually represent the covariance between an individual *input* ($X = p_i, q_i, r_i$) and the *response* (y_i) data,

$$y_i = y_0 + f_1(p_i) + f_2(q_i) + f_3(r_i) + \epsilon_i \quad (6)$$

The choice for fitting functions (f_j) is flexible, although a typical choice is a natural cubic spline. Natural cubic splines are a collection of polynomials, with second derivative equal to zero at the end points or knots. By specifying more knots, the splines can represent a higher frequency fluctuations. The fit between y and $f_1(p)$ can be generated through any penalized linear least-squares algorithm,

$$\left\|\boldsymbol{y} - \mathrm{f_j}(\boldsymbol{x})\right\| + \lambda \int_0^1 \left[\mathrm{f_j}''(\boldsymbol{x})\right]^2 dx = 0 \quad (7)$$

The fit penalization, λ, is the primary means by which the solution is tuned. The fit between *y* and the sum of f_j's means that the influence of each f_j on the global solution can be observed, plotted and evaluated. As mentioned, this is one of the principal strengths of the GAM and it permits a more interactive and nuanced approach to determining the significance of each input variable and the behavior of each f_j.

We implemented the penalized least squares using the Ridge regression algorithm in the Scikit-Learn library with a specified value for penalization and normalization of all input variables;

```
>> model = Ridge(alpha=λ,normalize=True).
```

The natural cubic spline matrix with k=9 knots was implemented using the Patsy module

```
>> basis = dmatrix("cr(train,df=10)-1", {"train": X[j]}).
```

We incorporated this penalized regression into the global fit using the Backfit Algorithm (Wood, 2017), which permits an iterative approach to fitting where each environmental correlate, j, is fit against the partial residuals (e_p), or the difference between the signal response (s) and the spline fit to all inputs except X_j,

$$e_p^j = \hat{s} - \sum_{k \neq j} f_k(X_k) \quad (8)$$

Here, s has already been standardized or normalized to have zero mean. The Backfit Algorithm described by Wood (2017), has been reproduced here for clarity. The Python code can be found in the Supplement

1. Standardize or remove the mean from s: $\hat{s} = s - \bar{s}$
2. Set the initial spline functions to zero: $\mathbf{f_j} = 0$
3. Use linear regression to fit f_j to e_p: `basis = model.fit(basis,ep);`
4. Estimate y from f_j: $\hat{s} = \sum_j \hat{f_j}$, `news = basis.predict(dmatrix("cr(valid, df=10)-1", {"valid": X[j]}))`
5. Recompute e_p: $e_p^j = \hat{s} - \sum_{k \neq j} \hat{f}_k(x_k)$,
6. Repeat steps 3 thru 5 until e_p stops changing.

More complex examples, involving other link functions between y and f_j, the imposition of different probability distributions on y_i (e.g. Gamma, Poisson or exponential) are all treated in more detail in Hastie et al., (2001).

To determine the optimal fit, we iteratively apply the Backfit Algorithm to the training data subset, and then compute the Generalized Cross Validation (GCV), as it varies with λ, the penalization parameter,

$$V(\lambda) = \frac{\frac{1}{n}\|I - \boldsymbol{A}(\lambda)y\|^2}{\left[\frac{1}{n}tr\left(I - \boldsymbol{A}(\lambda)\right)\right]^2} \quad (9)$$

In equation (9), n is the number of records of instrument signal response, I is the identity matrix, and A is the "influence" matrix, reflecting the penalized linear least-squares solution that can be applied as a step during the GAM fit (Golub et al., 1979),

$$\boldsymbol{A}(\lambda) = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + n\lambda I)^{-1}\boldsymbol{X}^T \quad (10)$$

The GCV approach is to look for the minimum in $V(\lambda)$ to determine the most appropriate regularization penalty and strike the best balance between fit complexity and over-fit. The GCV metric is better suited for this task than seeking the minimum residual sum of squares, because that value decreases continuously with n and with the magnitude of λ.

The GCV score can be computed directly using equation (9). It is also computed and can be output by the Sci-kit Learn Regression() toolbox. We used Ridge Regression, and the GCV score is output as,

```
>> model = Ridge(alphas=λ,store_cv_values=True).fit(X_train,s_train)
```

```
>> gcv = model.score(X_test,s_test)
```

Because the components of the GAM model are separable, it is also possible to determine which environmental correlates contribute most to the best-fit solution. This avoids the inclusion of correlates that make no contribution or may even degrade the GAM solution. The Bayesian Information Criterion (BIC) considers the model fit quality, but also penalizes for models of increasing complexity (Burnham and Anderson, 2004), providing a measure for each correlate's contribution to the GAM solution,

$$BIC = n\log_e(RSS/n) + k\log_e(n) \quad (11)$$

This version of the BIC applies when using a maximum likelihood estimator (such as ridge regression). The term *k* is the number of parameters included in the model. In this case, *k* is equivalent to the number of environmental correlates. The absolute value of BIC is not important; rather, the goal is to seek a minimum in BIC, which indicates the model best fit with the fewest parameters. For this task,

$$\Delta BIC_i = BIC_i - \min(\mathrm{BIC}) \quad (12)$$

will achieve a value of zero when the best set of environmental correlates have been used. The ΔBIC is further useful as it allows the user to determine if certain environmental correlates degrade the overall solution or make no contribution (Figure 6).

2.6 Implementation of the LSTM algorithm

Recurrent neural networks (RNN) can be used to interpret sequential data, like time series, where each data record may be related to the records that preceded it. The Neural Network uses functional dependencies along a network of nodes and the influence of these dependencies are weighted based on their relative importance. The RNN keeps track of these network weights as a means to archive predictive information as memory (Brownlee, 2019b). Since their development, RNNs sometimes have difficulty converging to a solution when attempting to optimize weights at all the nodes. This problem was solved by the Long Short-Term Memory algorithm (Hochreiter and Schmidhuber, 1997) that discards or "forgets" weight information that is not pertinent to the solution. The documentation of RNN theory, concepts, and implementation is very extensive, rapidly evolving and available in the public domain, so we will move straight to a discussion of the implementation for instrument bias correction. We used the Keras API (Chollet, 2018) which serves as an interface to the Tensorflow toolbox to develop, train, and implement the LSTM network.

2.6.1 Taxonomy of the time series forecast

Because there are so many types of problems that can be solved using Neural Networks, it is helpful to list out the characteristics of this particular time series solution, because this affects the structure of the neural network (Brownlee, 2019b). In our case, we are determining a *single* output from *multivariate* inputs; the neural network is a regression, rather than classification, we seek a *multi-time step* output to be able to predict over an unspecified range of time, and the current solution is static because it has been trained in-situ calibration data and does not update the solution over time. The exogenous inputs are water temperature and water hydrostatic pressure that the SWIMS experiences. The endogenous inputs, which

are co-influenced by the environment are water vapor inside the SWIMS detector, measured at m/z=18, the sample temperature, the circuit board temperature, and the mass spectrometer background noise measured at m/z = 5 **Error! Reference source not found.**(Figure 4).

Instrument bias correction can be thought of as time series prediction. Even though our approach is to use a multivariate set of inputs to help develop the bias prediction, the potential for long term transients in the instrument signal encourage the interpretation of bias correction as a sequential and time dependent statistical problem. Examples of instrumental memory can include, e.g. the silicon membrane stiffening (see Section 2.2), or the thermal inertia a pressure casing that may dampen the heat transfer between the environment and electronics inside the housing. We use the Keras Sequential() model. The 2D environmental array *X*, of *n* data records through time by *k* input parameters (e.g. temperature, pressure) must be reshaped into a 3D array or tensor. The *n* data records in time are decomposed into *p* sequences of *t* time steps: $n = p \times t$ (Stevens and Antiga, 2019). Tensor creation provides the RNN with multiple time series realizations against which to train and develop network weights. The fundamental choice for the user is to decide how many *t* time steps to include in each sequence. If data is periodic, it may be instructive to break the data into lengths that roughly capture an interval of the period. For example, two years of solar radiation data or sea level data measured every 10 minutes may be naturally broken into *t* = 144 or *t* = 36 time steps corresponding to one day or one half tidal period. However, this choice is rarely carried out a priori and must be determined iteratively.

After the $p \times t \times k$ tensor dimensions have been established, the user must choose a functional relationship or "activation function" between input and response at each network node, they number of iterations or "epochs" over which the RNN algorithm will train, the number of "neurons", and the 'optimizer" or metric that is used to evaluate the goodness of fit. As with the time steps, the settings for these parameters cannot be determine a priori, so we establish appropriate values through iteration (Brownlee, 2019b).

Keras allows a user to take control of when the RNN weights are updated; this is known as controlling the model state or "stateful=True". By default, Keras updates the LSTM state after a "batch" is processed. A batch is a collection of sample sequences, where each sample sequence has *t* timesteps, as we defined above. A batch size of 1 causes the model weights to be updated after each sample, but the penalty in processing speed and computation, often requires a large batch size. Ideally the batch size is a factor of *p*, the number of sample sequences, otherwise a set of left over sequences are processed in an additional step (Brownlee, 2019b).

2.7 Determining fit quality

During tuning and iteration of the GAM model, we used GCV(λ) to test for overfitting and the root mean square error (RMSE), which is a measure of the deviation between modeled bias $\hat{\beta}(X)$ and instrument bias, using the train data sets. We also evaluated neural network LSTM model using the RMSE between $\hat{\beta}(X)$ and the instrument bias, measured during in-situ calibrations.

To evaluate the overall fit quality, we measured the RMSE between the independent O_2 and CO_2 instruments (y_{ind}), and the bias corrected signal from the QMS and SWIMS instruments as defined by equation 5, ($y(\vec{x}, t)$):

$$RMSE = \sqrt{\frac{1}{n}(y(\vec{x}, t) - y_{ind})^2} \quad (13)$$

3.0 Results and Discussion

The bias correction workflow is depicted in Figure 1 Figure 7; the calibrated GAM solution has been graphed in Figure 7 panels b through d, but the steps are essentially the same for the LSTM solution. In this section we present the details of the GAM and LSTM fits and contrast the two bias correction models.

3.1 GAM fit:

The ability to choose a functional form for each X_j environmental correlate was an attractive feature of the GAM, because early tests revealed that oxygen (m/z=32) strongly correlated with water vapor (m/z=18), and signal from the SWIMS showed m/z=18 ion currents outside the range observed during in-situ calibration. Consequently, it appeared necessary to have a linear or proportional correction to m/z = 18. Water is present in solution at nearly 1 mol/mol, so its concentration far exceeds the other analytes. Somewhat counterintuitively, m/z=18 correlated positively with m/z=32, perhaps suggesting a similar response to membrane permeability, rather than competition for ionization inside the SWIMS source (Figure S1 in the supplement**Error! Reference source not found.**).

All the environmental correlates (Figure 5) negatively covaried with the water depth. More subtle features, such as lag between the circuit board temperature (uC Temp) and the sample temperature can also be observed in the SWIMS electronics temperature (Figure 5, panel b). Using the flexibility of the GAM, we tested both linear and quadratic fits between m/z=18 and the target output variable, [O_2] or m/z=32. While these parameterizations showed a stiffer, more proportional response to the large-scale variations in m/z=18, ultimately the natural cubic spline produced the best RMSE solution.

Having chosen a cubic spline functional form for *f()* for each X_i, there remain only two additional parameters that can be used to tune the solution – the number of knots in each spline and the value of the penalty function, λ (equation 9). We tested the fit to in-situ calibration data for a range of from 3 to 30 knots and observed no significant change in fit quality above 10 knots, so all cubic spline fits used a total of 10 knots. The term GCV(λ) was computed iteratively over a range of $\lambda = 10^{-10}$ to $\lambda = 10^{10}$ (Figure 8); the minimum GCV(λ) suggests the region where fit complexity and minimization of bias are optimal (Wood, 2017). We found GCV(λ) was not sensitive to the penalty, outside the range $10^{-2} < \lambda < 10^{5}$, with a minimum near $\lambda = 10^{5}$, so this value of the penalty was implemented in the solution.

3.2 LSTM fit:

As noted, the Keras LSTM algorithm requires iteration to choose appropriate values for the *t* time steps in each sample, the batch size, and epochs, as well as the choice for how often

to update the weights of the RNN, or statefulness. We chose to optimize based on the RMSE, used a hyperbolic tangent activation function. We found the LSTM solution was most sensitive to batch size and the number of epochs, especially as they related to overfitting. To mitigate overfitting, we implemented node dropout regularization using the Keras Dropout() attribute. The approach is to assign a dropout likelihood between 0 and 1, wherein the model will randomly remove some nodes during training, thereby reducing co-dependence and overweighting of certain nodes (Srivastava et al., 2014).

Because the choice of batch_size, epoch number, and dropout regularization cannot be determined a priori, but have a preponderant influence on overfitting, we objectively determined the optimal values for these three hyperparameters using the GridSearchCV() algorithm in Keras. The approach tries all permutations of the hyperparameters and measures the fit quality using the RMSE and a k-fold cross validation (with k=5). The k-fold cross validation randomly samples the training data to produce test data subsets, which are then used to measure fit quality k times. We tested batch_sizes ranging from 20 to 80, epoch numbers ranging from 5 to 30, and dropout likelihood ranging from 0.1 to 0.8. The smallest k-fold RMSE value was found at with a batch_size = 80, epochs = 20, and Dropouts = 0.4. The residual error between the training data and the LSTM solution - 'train RMSE' in Figure 9 reveal a continual reduction in both test and train RMSE through epoch = 20. Beyond epoch=20, the Test RMSE increases, suggesting an overfit (Figure 9).

Finally, the choice of *t* timesteps in each sample can be an important consideration. Because time series may have quasi-periodic correlations, it is desirable to have *t* be large enough to capture the full period, in order to make future predictions based on past time series behavior (Brownlee, 2019a). The Triaxus tow vehicle was programmed to ascend and descend at 0.2 m/s, so a full tow from surface to 150 m and back to the surface took approximately 25 minutes, or $t = 750$ time steps at data reported every 2 seconds, which is the scan rate of the SWIMS. Initially, we anticipated that a sample size of $t > 750$ would provide the best fit. However, splitting the in-situ calibration data into a test and train subset did not permit the inclusion of sample sizes of $t = 750$, because we felt it necessary to validate against a test data set that was at least $2t$ in length. In practice, we tested values of $t = 50$, 100, 200 and 300. The Test RMSE actually improved significantly as *t* was reduced. Eventually, $t = 100$ provided both computational efficiency and low RMSE, even though this number of time steps does not encompass the full profile tow. The tow profile may not be as necessary, suggesting that the information used to reconstruct the bias, comes from the environmental correlates that are available at the prediction timestep, rather than from the learned temporal dependence.

3.3 GAM vs. LSTM bias correction, SWIMS tow

Normally, the procedure to evaluate a statistical learning algorithm involves validating the solution against the test data (see Section 2.4), which was set aside before the training stage. However, the independent measurement of oxygen by the SBE43 (described in Section 2.1) provides an opportunity to quantify the bias correction against an entirely unique measure of oxygen. It should be noted that the SBE43 probe can also be subject to its own sources of bias, some of which may not be accounted for, but this instrument has a long performance

history in oceanography (e.g. Helm et al., 2011) that support the choice to use it as a reference instrument.

The final list of environmental correlates was determined using the ΔBIC metric (equation 12). In addition to water vapor, we tested for environmental covariation in the water pressure, seawater temperature, the sample temperature inside the SWIMS heater block, the circuit board temperature, the temperature of the turbo pump, current draw of the turbo pump, and the duty cycle of the membrane heater. Using the ΔBIC it was determined that these last parameters did not add any meaningful additional constraints beyond what the first six environmental correlates. That is, the ΔBIC achieved a minimum after including water vapor, water pressure, circuit board temperature, sample temperature and instrument noise at m/z = 5 (Figure 6). The remaining correlates were eliminated from the GAM solution.

The SWIMS tow between 35° and 40° N, recorded a total of N = 49,181 individual measurements of dissolved O_2. A contour plot of dissolved O_2 reveals the tracer field in Figure 10. The RMSE between SBE43 and bias-corrected SWIMS data using the GAM was 11.2 μM (micromoles per liter of seawater); the units of RMSE are the same as the concentration data itself. The mean [O_2] in this section was 196 μM, suggesting a 5.7% deviation between the two instruments. Within the same section, the neural network LSTM bias correction yielded RMSE = 9.8 μM or 5.0% deviation overall. Both GAM and LSTM bias corrections tended to fit some regions better than others, however the fit quality of the GAM and fit quality of the LSTM did not degrade in the same places, suggesting some differences in how the two models respond to the environmental correlates (Figure 5).

It should be noted that we are focusing on interpretation of the relative RMSE between the GAM and LSTM solutions. The absolute value of the RMSE is less meaningful, because the calibration intercept (s^0) was not measured on the SWIMS in-situ calibrations. This term, s^0, represents the instrument baseline drift, and so we determined s^0 by optimal fit to the SBE43. The same baseline drift can be determined by fitting to another independent reference, such as the equilibrium oxygen solubility (Garcia and Gordon, 1992). When we use equilibrium solubility, the shape or trend in the daily estimates of s^0 remains the same, but the magnitude of s^0 shifts, causing a larger misfit between the SBE43 and the SWIMS. During future in-situ calibrations, we think it is possible to implement a workable measure of s^0 by shutting off the water pump, causing all the gas around the silicon membrane to be depleted and achieving a practical value of zero concentration for all gases except water vapor.

3.4 GAM vs. LSTM bias correction, Shipboard QMS

The bias corrections in the shipboard QMS were fit using training data over a four-day period of the surface ocean equilibrator time series from May 27 to June 1. The RMSE between the GAM solution and training data subset was 3.5%, and the LSTM misfit was 1.8%. Unlike the SWIMS tows, it was not possible to evaluate $\beta(X)$ independent of the environmental signal; $y(\vec{x}, t)$. The daily calibrations with reference gases did not take place for long enough to properly observe and decompose the time series aliasing. Instead, it was necessary to train the LSTM and GAM models on a section of the real time series. This approach can lead to muddling the separation between $y(\vec{x}, t)$ and $\beta(X)$, potentially correcting away some of the environmental signal in pCO_2 during the bias correction. However, the ambient changes in

564 pCO_2 should reflect the biology and chemistry which in turn are only partly dependent on the
565 exogenous environmental correlates. The endogenous environmental correlates reflect
566 instrument behavior, which should have zero correlation with environmental pCO_2. The
567 environmental correlates used to develop the bias correction model included, (1) temperature of
568 the lab where the QMS was installed, (2) the total gas pressure in the QMS measured as
569 voltage, (3) the seawater flow rate through the turbulent equilibrator, (4) water vapor measured
570 at m/z=18, and (5) m/z=15. Similar to the SWIMS tow, we found that three environmental
571 correlates caused an increase (no decrease) in the ΔBIC metric, signaling that they contributed
572 no meaningful constraint. Consequently, the IR pCO_2 cell temperature, the water wall flow rate,
573 and the second equilibrator temperature reading were eliminated from the bias correction
574 solutions (Figure 6).
575 After bias correction, the raw ion current was calibrated to CO_2 partial pressure, using
576 the three-point calibration of reference standards that were measured daily. There are
577 additional corrections to gas measurements that are made using a turbulent equilibrator and
578 these are described by Takahashi et al.,(2009). These corrections have not been implemented
579 here; while their implementation might improve the overall misfit between the two
580 measurements of pCO_2, they would drop out of the comparison between GAM and LSTM bias
581 corrections, so these additional data corrections are not material to this evaluation.
582 In this case, the GAM model was better at removing the periodic oscillation in the QMS
583 ion current at m/z=44 (**Error! Reference source not found.**). However, a level of noise
584 persists even after the bias correction, suggesting that the environmental correlates may be
585 missing some component of the bias. In total, the 18 day time series contains 5043 unique
586 measurements of pCO_2 by infrared absorption spectroscopy and by QMS. The RMSE between
587 the IR pCO_2 and GAM-corrected pCO_2 was 31.3 μatm; the average pCO_2 was 411 μatm,
588 revealing an overall misfit of 7.5% (Figure 11). The LSTM RMSE was 35.2 μatm or 8.5% of the
589 mean pCO_2. In this case, it appears the LSTM (not pictured) may have slightly overfit the
590 training data, resulting in a degraded fit to the overall time series. Nevertheless, the difference in
591 RMSE between GAM and LSTM was less than 1%, which suggests that both methods produce
592 very similar overall bias correction outcomes.
593
594 **4.0 Summary**
595 This study presents two models for instrument bias correction, a Generalized Additive
596 Model (GAM) and a Long-Short Term Memory (LSTM) Neural Network model. The two models
597 represent philosophically different approaches to the multivariate prediction; the GAM allows the
598 user to investigate the intermediate model fit products and choose the functional form *f()* for
599 optimal regression between the results and the individual environmental correlates in *X*. This
600 advantage was particularly useful when interrogating which environmental influences to include
601 as correlates in the model solution, using the ΔBIC criterion. This calculation is straightforward
602 and can be determined offline without iteration of the GAM model, precisely because the
603 solution is separable. The procedure eliminated three environmental correlates from both
604 Shipboard and SWIMS ocean data sets (Figure 6). The six remaining correlates were also used
605 to fit the LSTM solution.
606 The LSTM RNN model gives the user fewer intermediate diagnostics, which produces
607 an initial lack of confidence in the robustness of the solution, because it can be challenging to

understand or visualize the nature of the solution. Nevertheless, there is an emerging recognition that, compared to the human brain, computers are much more capable instruments at assigning appropriate weights to an n-dimensional set of variates in pursuit of a solution. By accepting these models, we implicitly acknowledge that the multivariate weights in the solution are beyond our capacity to evaluate simultaneously, thus rendering the “black box” criticism somewhat moot. However, the procedures for implementing RNNs, including the grid search or random search (B. Nakisa et al., 2018; Bergstra and Bengio, 2012), provide a systematic approach to determining the optimal tuning of hyperparameters (e.g. times steps, batch size, epochs, hidden nodes), and the eventual robustness of the solution has held up under rigorous testing and comparison. In the SWIMS data set with in-situ calibration, the LSTM solution proved more effective at removing bias in the high gradient oceanic region, with tows across the Gulf Stream. However, the GAM exhibited better fit quality in the Antarctic shipboard QMS data set, ascompared to the LSTM.

The difference between GAM and LSTM RMSE was 1% or less for both ocean sections, suggesting that both models performed similarly well. The RMSE for both methods were better than 6% for O_2 and less than 9% for CO_2, demonstrating a predictive accuracy of better than 91% for both dissolved gases. The quality of the bias removal solution was significantly more dependent on the availability of coincidently sampled environmental correlates as inputs. We further found that the in-situ calibration for SWIMS data was a significant factor in producing a high fidelity bias correction. Several attempts were made to produce the same bias correction using just SWIMS tow data (without the in-situ calibration) as training data, and the solution was significantly diminished with an RMSE for the LSTM model of 17% as compared to 5% with the in-situ calibration. These results demonstrate that the bias corrections are most effective when they can be tuned using the in-situ calibration with an invariant reference gas to reveal the instrument bias.

The overall performance of the GAM and LSTM models were highly comparable, making it difficult to declare a clear winner in this case. The primary advantage conferred by the GAM model is the ability to evaluate the fit to each individual correlate, separately. This is a big advantage when it is necessary to better understand an instruments behavior and might even lead to engineering solutions that eliminate the biggest source of bias. In comparison, the skill that an LSTM RNN brings to time series prediction can potentially serve to model longer term transients in the signal, which could lead to a better bias model when few or no environmental correlates have been measured.

Acknowledgements:

This research was supported by an award from the National Science Foundation Chemical and Biological Oceanography Program #1429940. We thank two anonymous reviewers for the comments and suggestions that have improved this manuscript. The GAM backfit algorithm is available at https://github.com/bloose/Python_GAM_Backfit. The supplemental contains annotated Python scripts and SWIMS example data to demonstrate application of the GAM and LSTM to bias correction.

References

Ahmed, N.K., Atiya, A.F., Gayar, N.E., and El-Shishiny, H. (2010). An Empirical Comparison of Machine Learning Models for Time Series Forecasting. Econom. Rev. *29*, 594–621.

B. Nakisa, M. N. Rastgoo, A. Rakotonirainy, F. Maire, and V. Chandran (2018). Long Short Term Memory Hyperparameter Optimization for a Neural Network Based Emotion Recognition Framework. IEEE Access *6*, 49325–49338.

Bell, R.J., Short, R.T., van Amerom, F.H.W., and Byrne, R.H. (2007). Calibration of an In Situ Membrane Inlet Mass Spectrometer for Measurements of Dissolved Gases and Volatile Organics in Seawater. Environ. Sci. Technol. *41*, 8123–8128.

Bergstra, J., and Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. J. Mach. Learn. Res. *13*, 281–305.

Brownlee, J. (2019a). Deep Learning for Time series Forecasting: Predict the future with MLPs, CNNs and LSTMs in Python (Jason Brownlee).

Brownlee, J. (2019b). Long Short-Term Memory Networks With Python (Australia: Jason Brownlee).

Burnham, K.P., and Anderson, D.R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. Sociol. Methods Res. *33*, 261–304.

Cassar, N., Barnett, B.A., Bender, M.L., Kaiser, J., Hamme, R.C., and Tilbrook, B. (2009). Continuous High-Frequency Dissolved O2/Ar Measurements by Equilibrator Inlet Mass Spectrometry. Anal. Chem. *81*, 1855–1864.

Chapman, D.C., and Beardsley, R.C. (1989). On the Origin of Shelf Water in the Middle Atlantic Bight. J. Phys. Oceanogr. *19*, 384–391.

Chollet, F. (2018). Deep Learning with Python (Shelter Island, NY: Manning Publications).

Dawson, P.H., and Herzog, R.F. (1995). Quadrupole mass spectrometry and its applications (New York: American Institute of Physics).

Delle Monache, L., Nipen, T., Deng, X., Zhou, Y., and Stull, R. (2006). Ozone ensemble forecasts: 2. A Kalman filter predictor bias correction. J. Geophys. Res. Atmospheres *111*.

Futó, I., and Degn, H. (1994). Effect of sample pressure on membrane inlet mass spectrometry. Anal. Chim. Acta *294*, 177–184.

Garcia, H.E., and Gordon, L.I. (1992). Oxygen solubility in seawater: better fitting equations. Limnol Ocean. *37*, 1307–1312.

Golub, G.H., Heath, M., and Wahba, G. (1979). Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. Technometrics *21*, 215–223.

Guegen, C., and Tortell, P.D. (2008). High-resolution measurement of Southern Ocean CO2 and O2/Ar by membrane inlet mass spectrometry. Mar. Chem. *108*, 184–194.

Hastie, T., Tibshirani, T., and Friedman, J. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction.

Helm, K.P., Bindoff, N.L., and Church, J.A. (2011). Observed decreases in oxygen content of the global ocean. Geophys. Res. Lett. *38*.

Hochreiter, S., and Schmidhuber, J. (1997). Long Short-Term Memory. Neural Comput. *9*, 1735–1780.

Jenkins, W.J. (1982). Oxygen utilization rates in North Atlantic subtropical gyre and primary production in oligotrophic systems. Nature *300*, 246–248.

Lee, W.S., Yeo, K.S., Andriyana, A., Shee, Y.G., and Mahamd Adikan, F.R. (2016). Effect of cyclic compression and curing agent concentration on the stabilization of mechanical properties of PDMS elastomer. Mater. Des. *96*, 470–475.

Newman, M.C. (1993). Regression analysis of log-transformed data: Statistical bias and its correction. Environ. Toxicol. Chem. Int. J. *12*, 1129–1133.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

Pimentel, K.D. (1975). Toward a mathematical theory of environmental monitoring: the infrequent sampling problem (United States).

Saltzman, E.S., De Bruyn, W.J., Lawler, M.J., Marandino, C.A., and McCormick, C.A. (2009). A chemical ionization mass spectrometer for continuous underway shipboard analysis of dimethylsulfide in near-surface seawater. Ocean Sci *5*, 537–546.

Short, R.T., Fries, D.P., Kerr, M.L., Lembke, C.E., Toler, S.K., and Byrne, R.H. (2001). Underwater mass spectrometers for in situ chemical analysis of the hydrosphere. J. Am. Soc. Mass Spectrom. *12*, 676–682.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J. Mach. Learn. Res. *15*, 1929–1958.

Stevens, E., and Antiga, L. (2019). Deep Learning with PyTorch (Shelter Island, NY: Manning Publications).

Sun, F., Xiong, R., and He, H. (2016). A systematic state-of-charge estimation framework for multi-cell battery pack in electric vehicles using bias correction technique. Appl. Energy *162*, 1399–1409.

Takahashi, T. (1961). Carbon dioxide in the atmosphere and in Atlantic Ocean water. J. Geophys. Res. 1896-1977 *66*, 477–494.

Takahashi, T., Sutherland, S.C., Sweeney, C., Poisson, A., Metzl, N., Tilbrook, B., Bates, N., Wanninkhof, R., Feely, R.A., Sabine, C., et al. (2002). Global sea–air CO2 flux based on

724 climatological surface ocean pCO_2, and seasonal biological and temperature effects. Deep-Sea
725 Res. Part II *49*, 1601–1622.

726 Takahashi, T., Sutherland, S.C., Wanninkhof, R., Sweeney, C., Feely, R.A., Chipman, D.W.,
727 Hales, B., Friederich, G., Chavez, F., Sabine, C., et al. (2009). Climatological Mean and
728 Decadal Change in Surface Ocean pCO_2, and Net Sea-air CO_2 Flux over the Global Oceans.
729 Deep-Sea Res. Part II *56*, 554–577.

730 Wenner, P.G., Bell, R.J., van Amerom, F.H.W., Toler, S.K., Edkins, J.E., Hall, M.L., Koehn, K.,
731 Short, R.T., and Byrne, R.H. (2004). Environmental chemical mapping using an underwater
732 mass spectrometer. TrAC Trends Anal. Chem. *23*, 288–295.

733 Wood, S. (2017). Generalized Additive Models: an introduction with R (CRC Press).

734

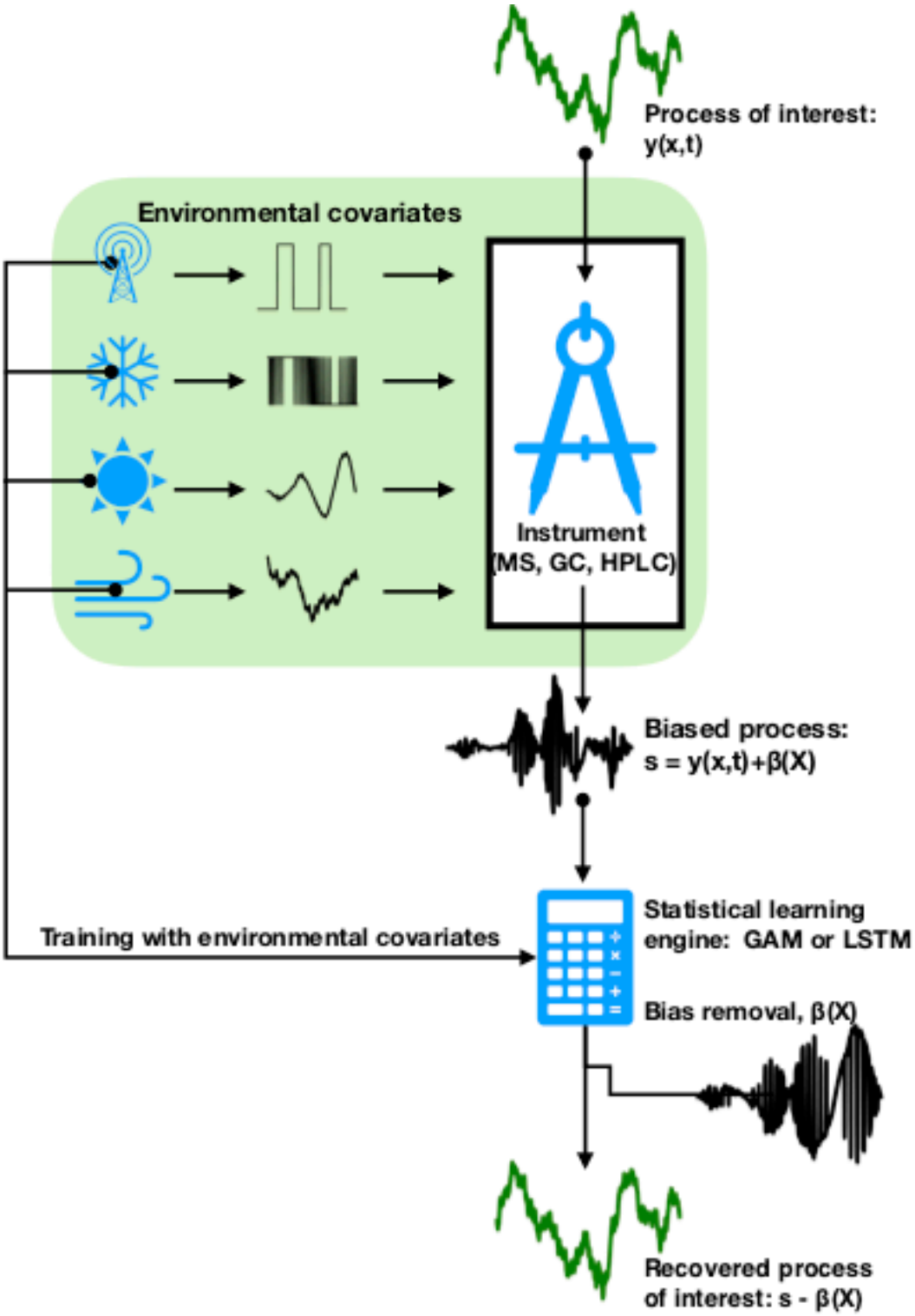


Figure 1. A schematic depiction of the effect of environmental factors on the introduction of bias into in-situ chemical instrumentation, and the subsequent identification and removal of bias using environmental covariates to train a statistical learning engine.

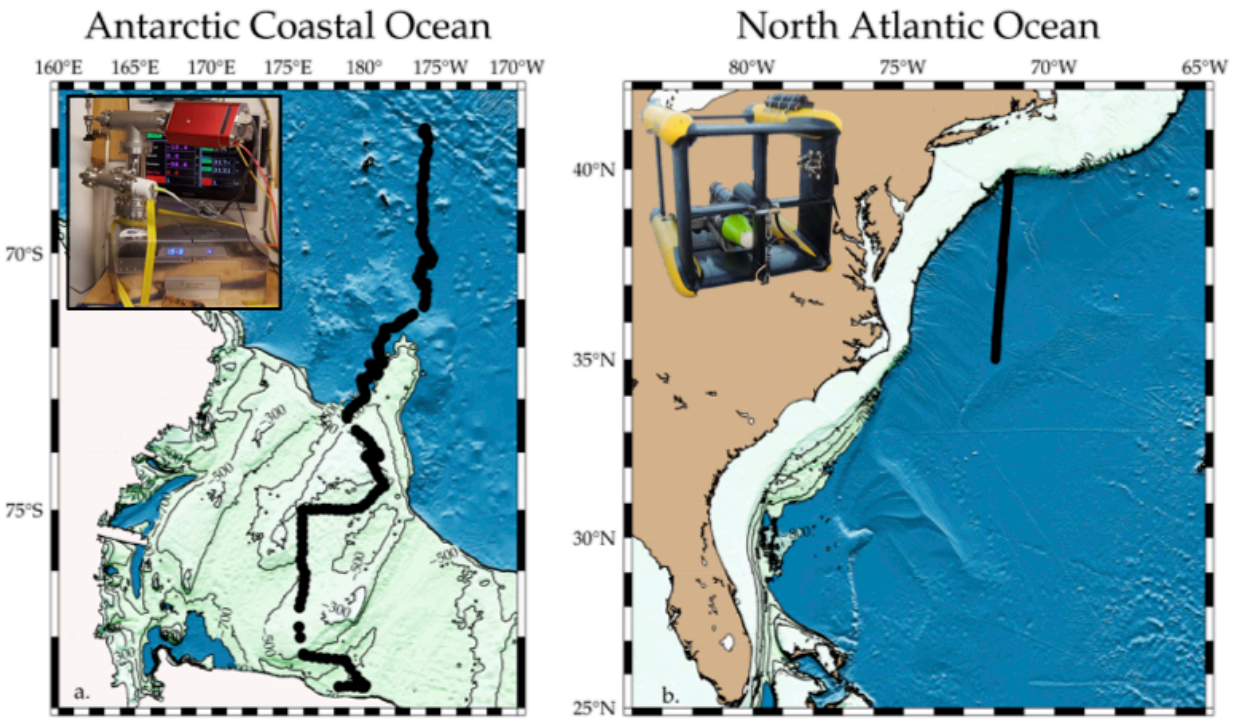
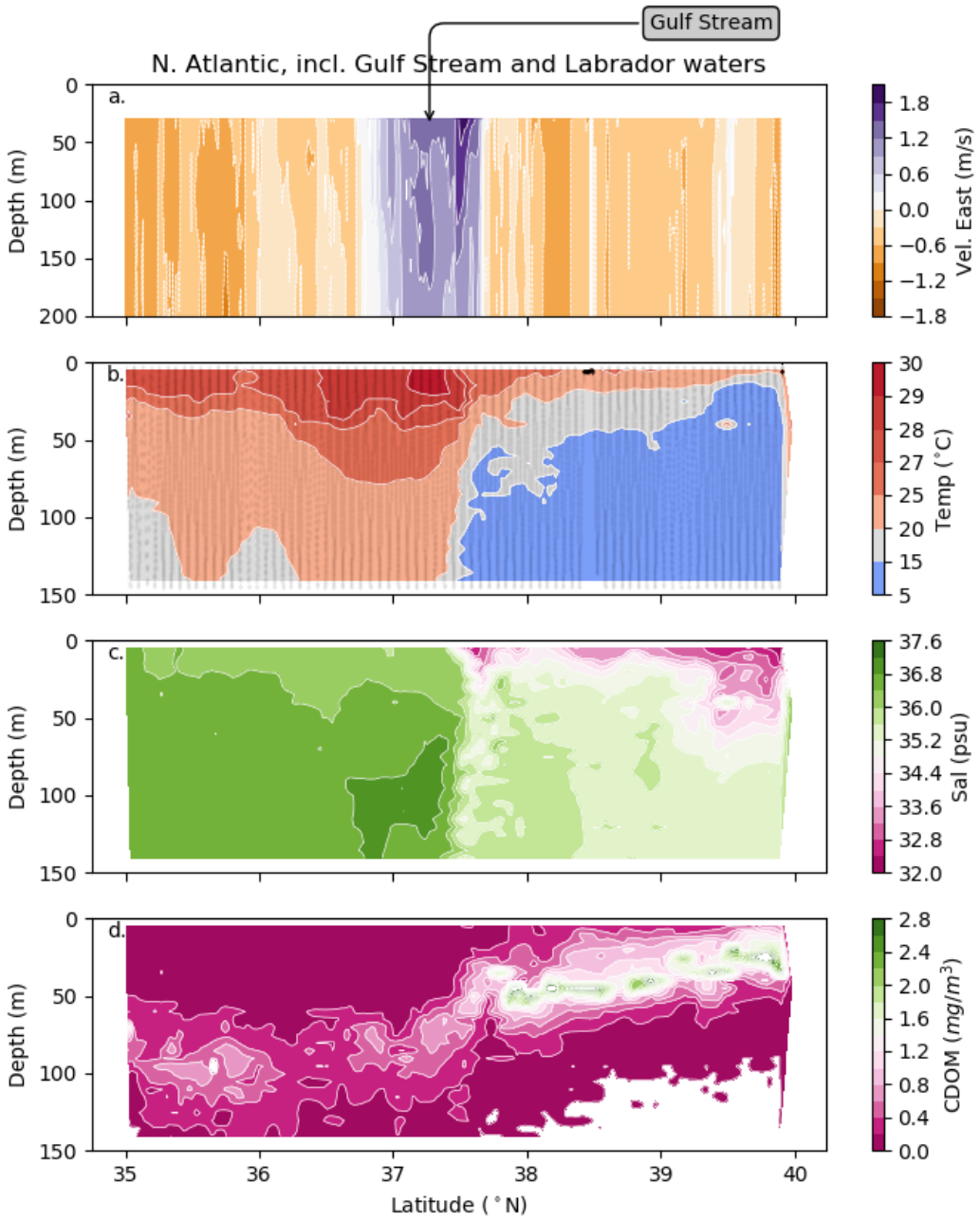


Figure 2. Maps showing the ocean regions where QMS data was collected as part of oceanographic surveys. Panel a shows a region of the North Atlantic, including the Gulf Stream and Labrador water; panel b is a region in the Ross Sea, along the coast of Antarctica in the Atlantic sector. The measurements collected at these locations are the subject of the Additive Model and Neural Network bias correction algorithms that we compare in this study.



744
745 *Figure 3. Ocean properties during the SWIMS tow in the N. Atlantic, across the Gulf Stream and into coastal waters*
746 *influenced by the Labrador Current. The track lines of the tow are shown in panel b.*

747

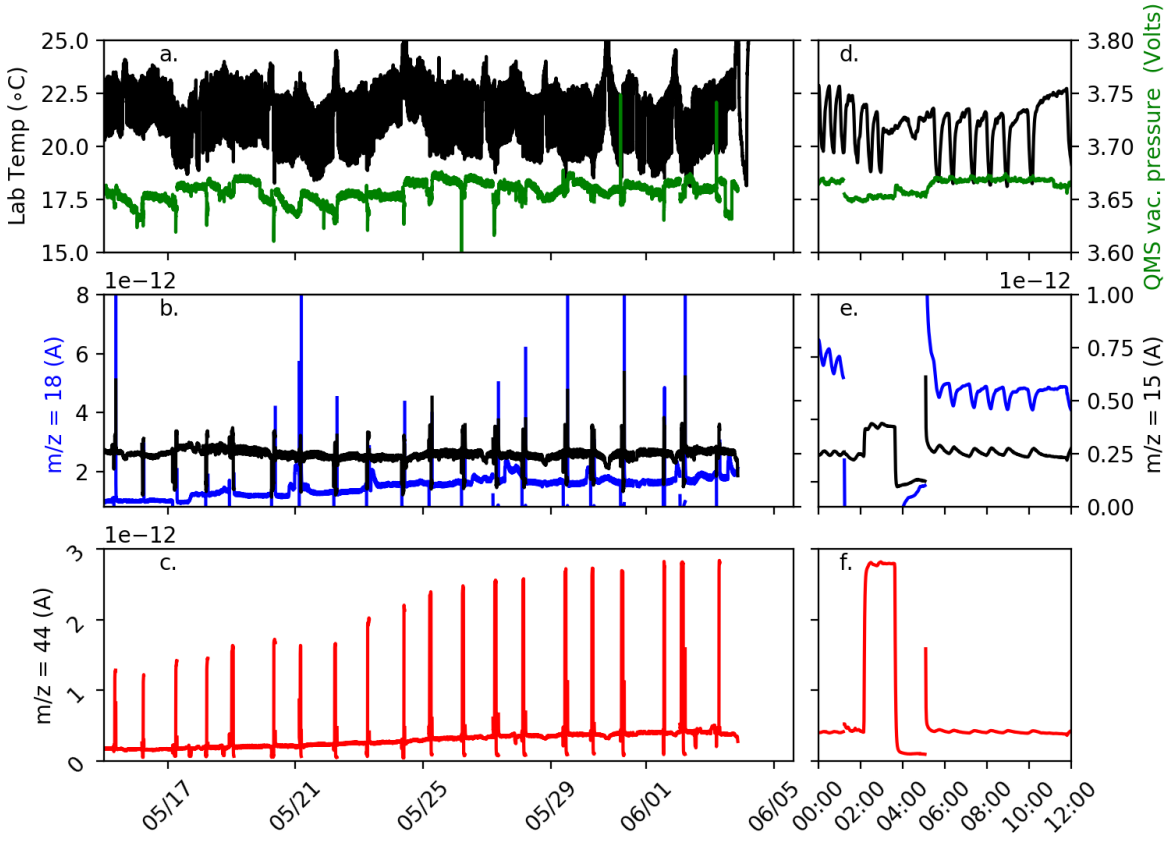


Figure 4. Time series of environmental correlates used to bias correct the pCO_2 signal measured by shipboard QMS at m/z = 44 (panels c and f). Panels a and d show the lab temperature and QMS vacuum pressure, panels b and e show water vapor (m/z=18) and m/z=15. Panels d e and f show the time series for one 12 hour period on June 1, 2017.

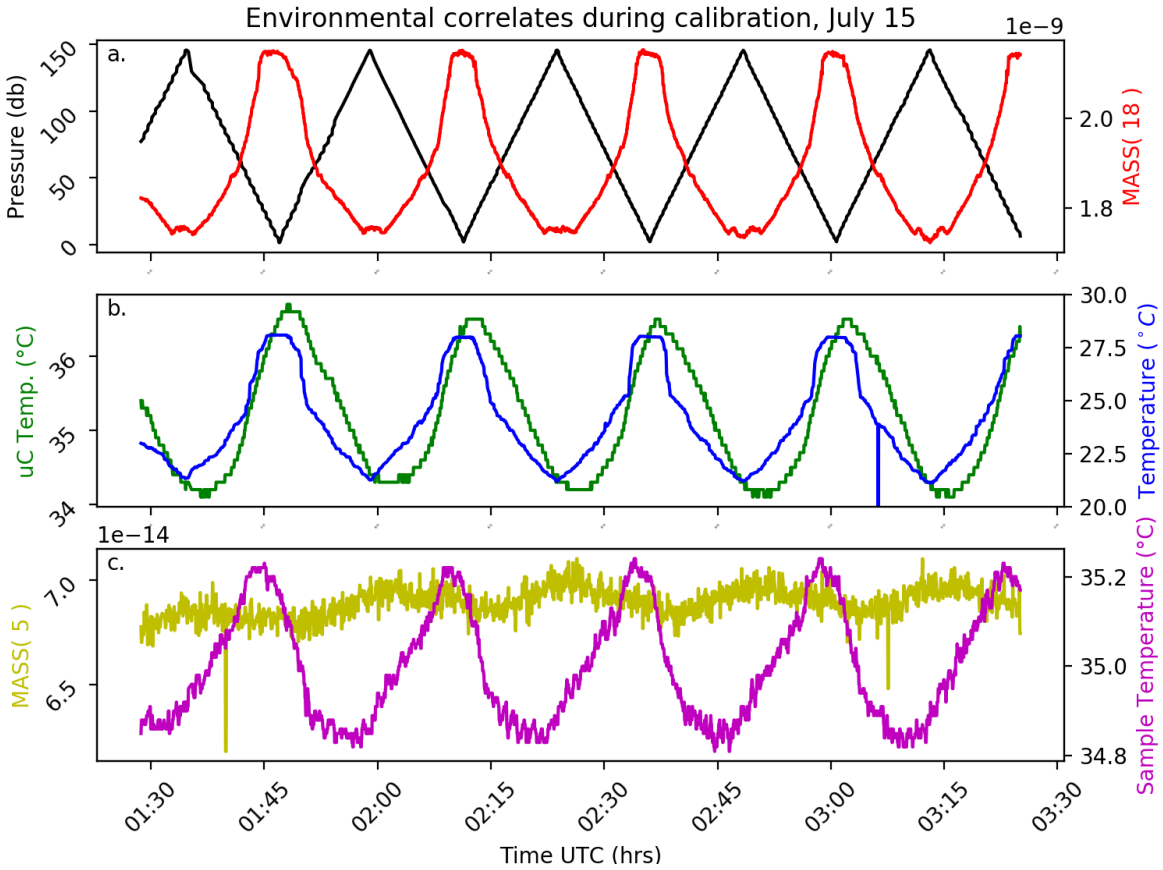


Figure 5. The six environmental correlates that were measured by the mass spectrometer or CTD to capture variations in the environment that are likely to influence the signal response (s) of the SWIMS for m/z=32 and other dissolved gases. Panel a shows instrument water depth as hydrostatic pressure and the ion current for water vapor (mass 18). Panel b shows the temperature of the electronics (uC temp) and the water temperature. Panel c shows the sample temperature or temperature of the heater block where gas is introduced across the membrane, and it shows ion current at mass 5, the electronic noise baseline.

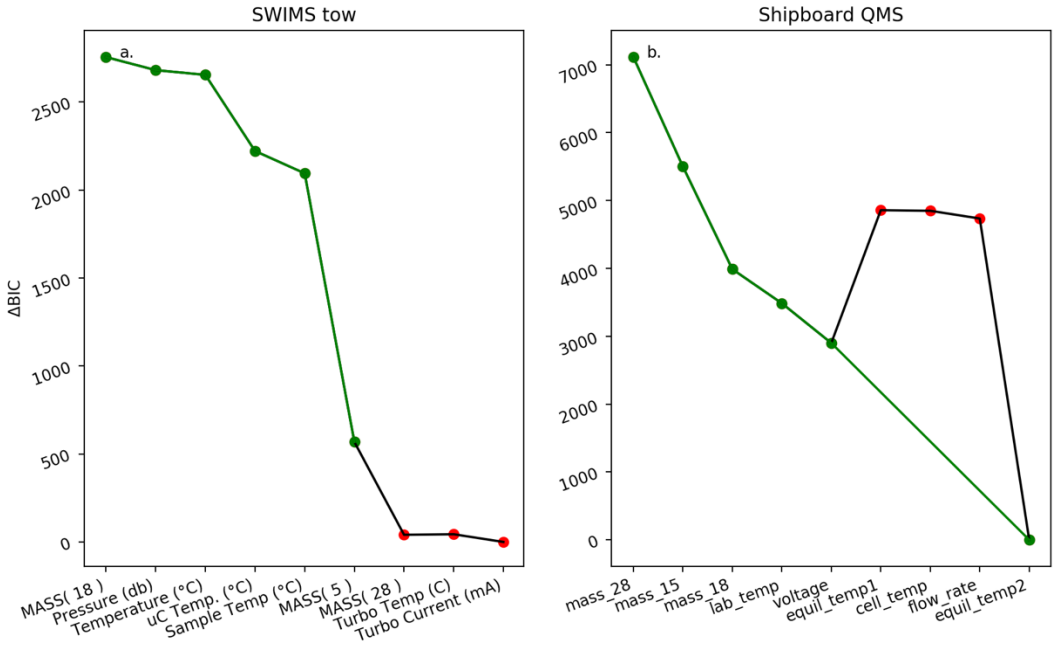


Figure 6. The normalized Bayesian Information Criterion (ΔBIC), which were used to determine the set of the environmental correlates that best reproduce the bias in the GAM model. The black line with red dots indicate environmental correlates that did not measurably improve the ΔBIC score. These were left out of the final set of environmental correlates, which are indicated by the green line and green dots.

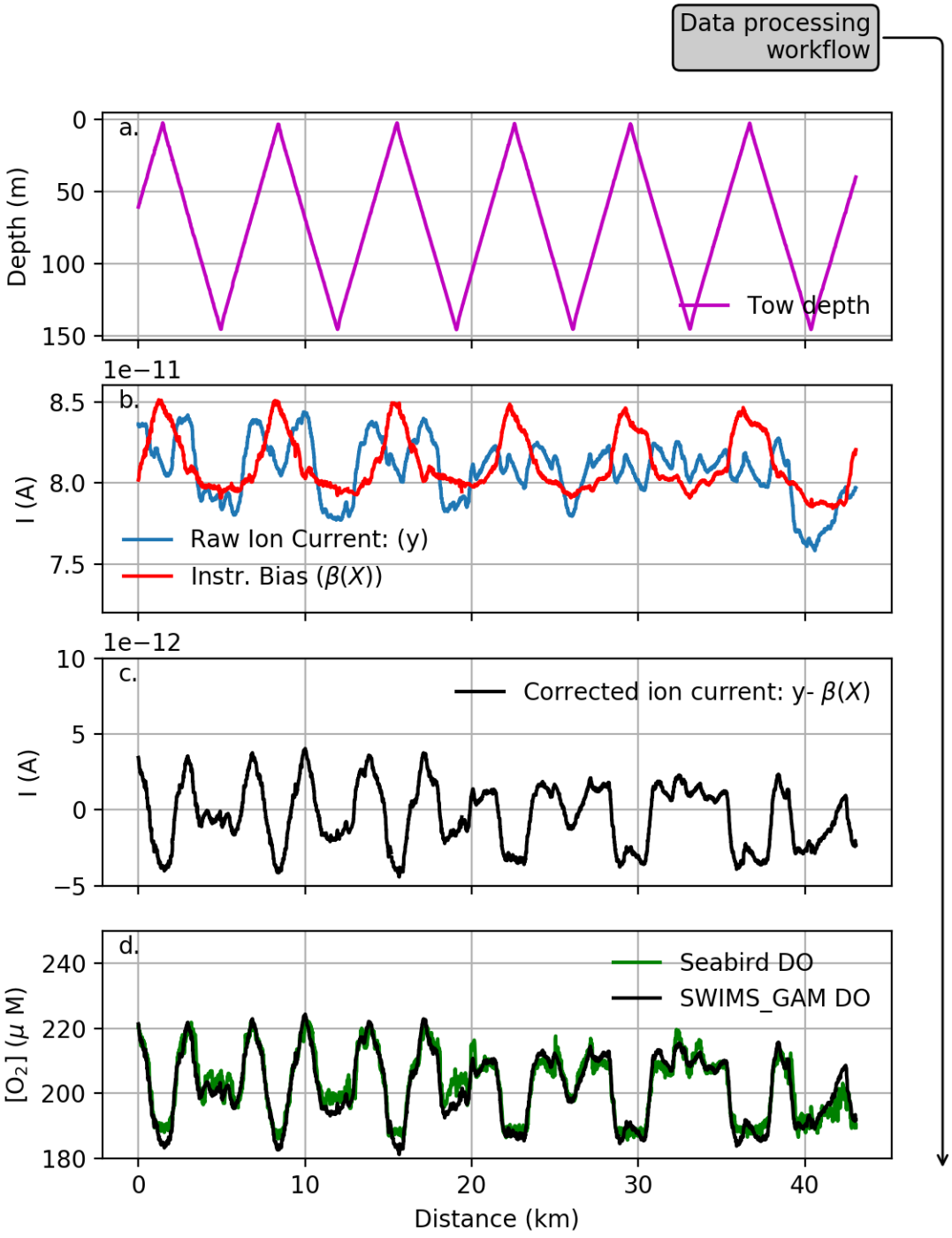


Figure 7. Sequence showing the SWIMS tow and bias correction using the Generalized Additive Model (GAM): Panel a reveals the depth recorded by the Triaxus CTD during vertical tows, panel b shows the raw signal recorded for oxygen at m/z=32, and the estimate of instrumental bias. Panel c shows the corrected ion current and panel d shows the calibrated ion current in O_2 concentration units, alongside the Seabird DO sensor.

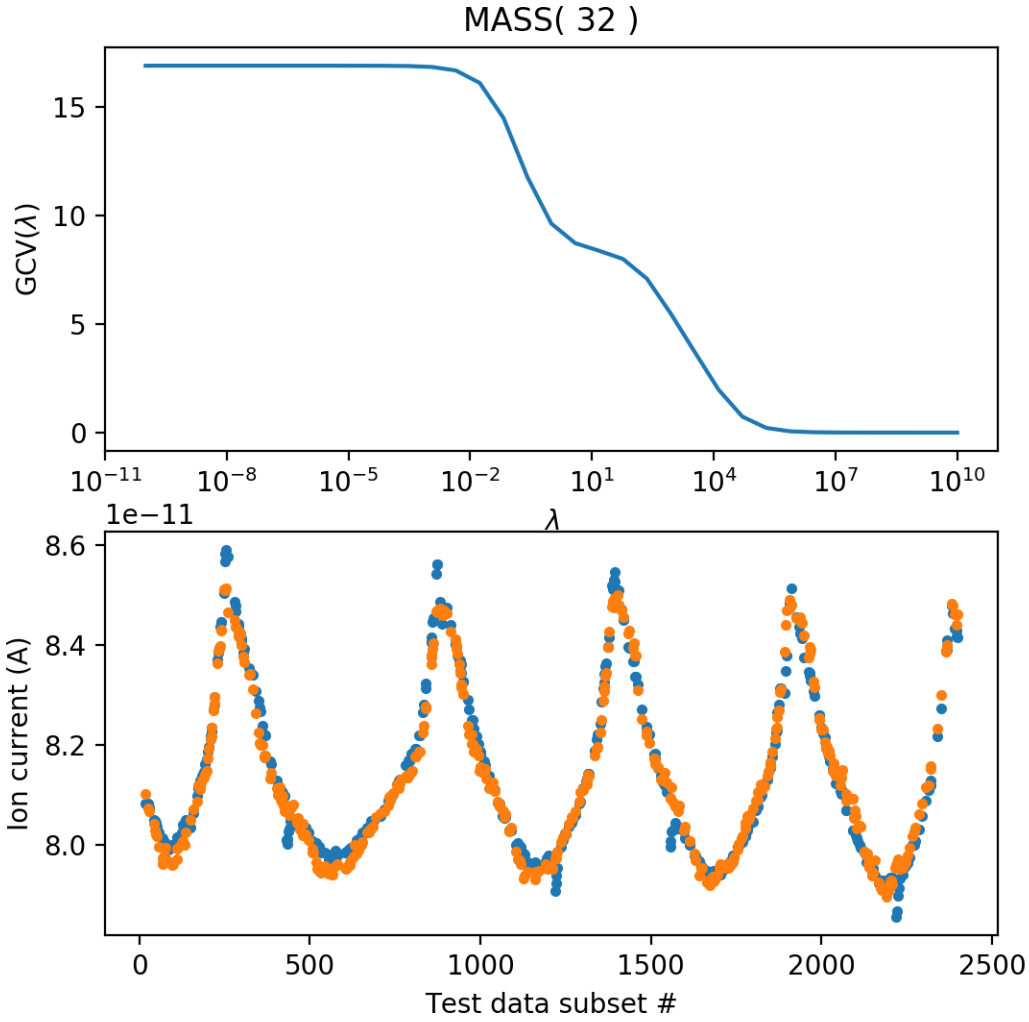


Figure 8: Test of the GAM solution using a range of values for the penalty term λ in equation 9.

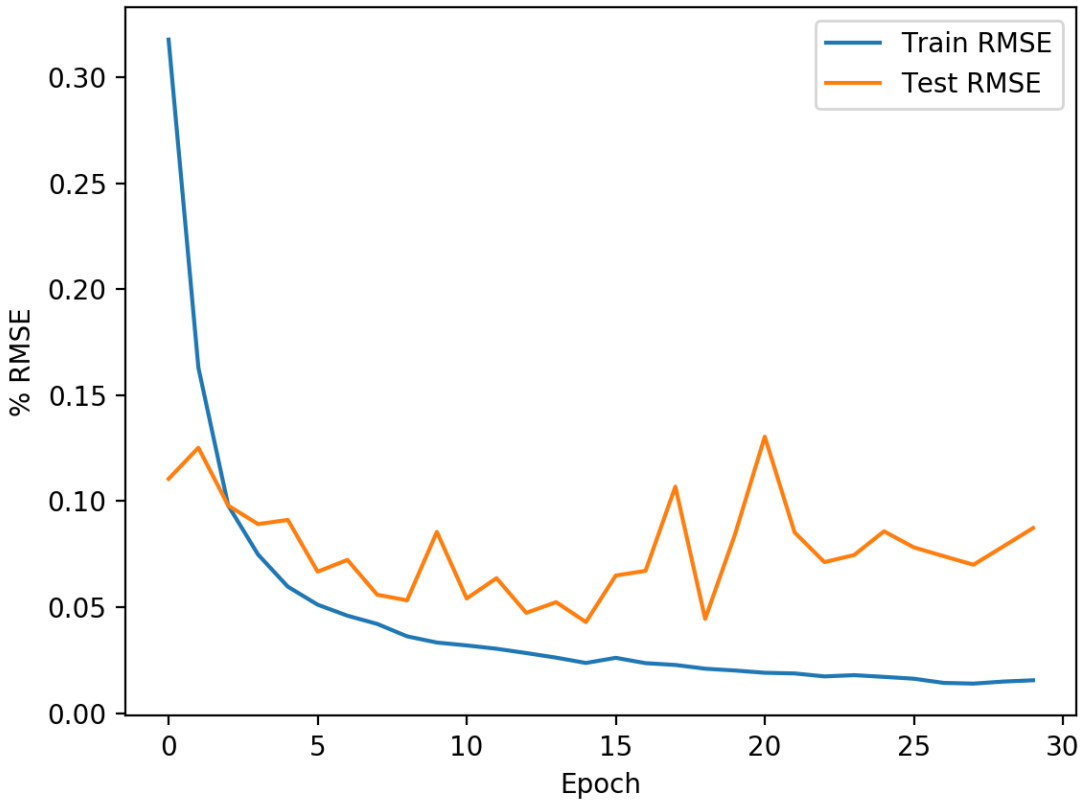


Figure 9. Root-mean squared deviation between the train and test subsets during successive training Epochs. While the training RMSE continually decreases, suggesting improvement, the Test RMSE begins to increase after 20 Epochs, suggesting that the solution is being overfit.

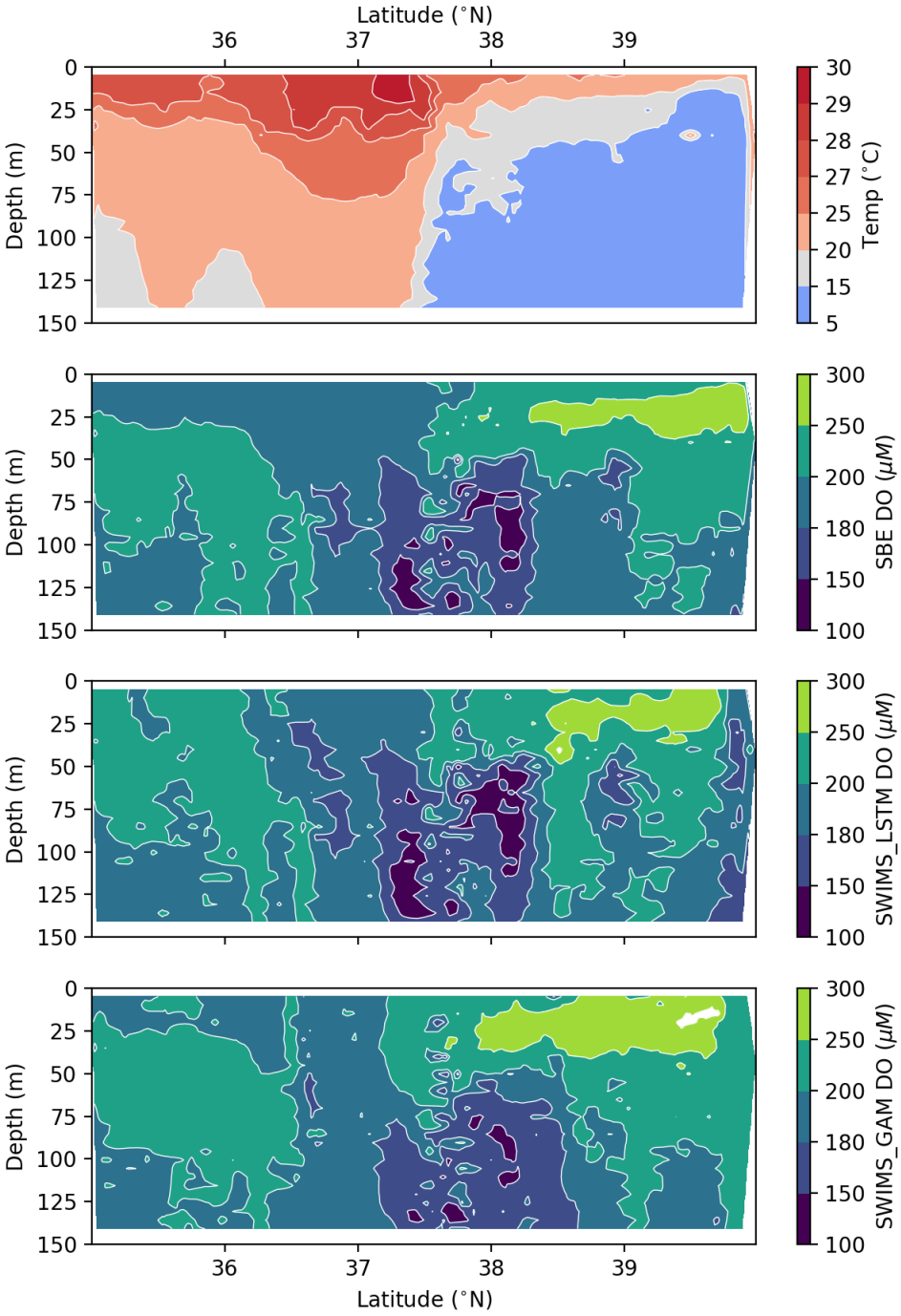


Figure 10. North Atlantic section, including Gulf Stream and Labrador waters showing temperature (top) and oxygen from the Seabird SBE 43 membrane and the SWIMS with bias corrections using GAMs and the neural network LSTM model.

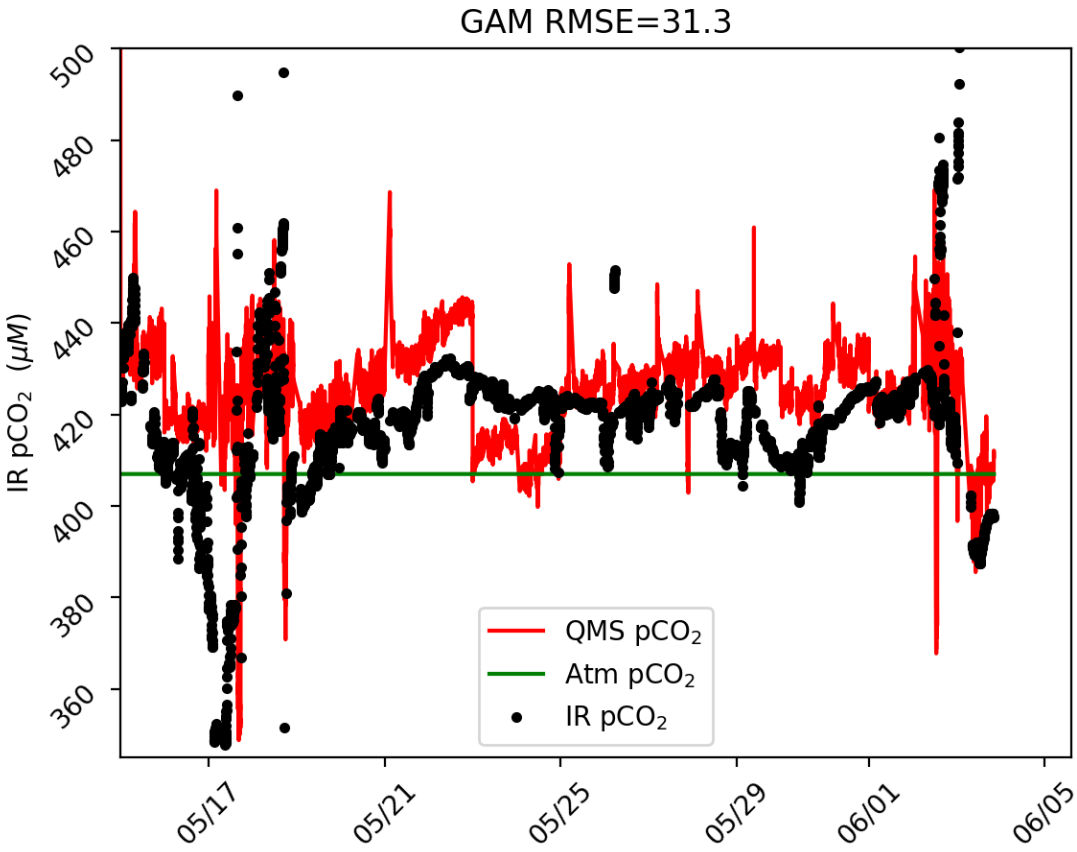


Figure 11. Bias-corrected and calibrated pCO_2 from shipboard QMS alongside measurements of pCO_2 by infrared absorption spectroscopy (IR pCO_2) in the Ross Sea, 2017.