Reviewer 2 Comments:

Q1:

This study presents two models for instrument bias correction, a Generalized Additive Model (GAM) and a Long-Short Term Memory (LSTM) Neural Network mode.

Biases in measurements from QMS and SWIMS measuring the dissolved CO2 and O2 in seawater, respectively, were estimated using the two models mentioned above. Corrected measurements from QMS and SWIMS were then compared to their reference measurements (IR spectroscopy and SBE 43, respectively) by performing simple linear regressions (calibration). RMSEs between the calibrated measurements and reference measurements were used to assess the models' performance.

Q2:

Advantages:
1. Both GAM and LSTM demonstrate a high degree of powerl at correcting for instrument bias using correlated environmental measurements.

Limitations:
1. The hyper parameters for the LSTM were determined in a quasi-empirical approach. The result may not be optimal.
2. The reviewer could not identify the details of how to separate the data set into training and testing sets.
3. K-fold cross-validation should be adopted when training and testing the models to have more comprehensive results.
4. In Figure 9, the LSTM began overfitting when epoch number was above 10; efforts such as reducing the network complexity should be taken in order to prevent overfitting instead of simply stop at epoch 5.
5. The review suggested a comparison base line such as simple multi-variable regression model should be included.
6. Lacking a comprehensive literature review.

Q4:

Is the English language of sufficient quality?
- Yes

Is the quality of the figures and tables satisfactory?
- Yes

Does the reference list cover the relevant literature adequately and in an unbiased manner?
- Yes

Are the statistical methods valid and correctly applied? (e.g. sample size, choice of test)
- No

Are the methods sufficiently documented to allow others to apply them?
- Yes

Is there adequate validation of the proposed method?
- No

Are the data underlying the study available in either the article, supplement, or deposited in a repository? (Sequence/expression data, protein/molecule characterizations, annotations, and taxonomy data are required to be deposited in public repositories prior to publication)
- Yes

Does the study adhere to ethical standards including ethics committee approval and consent procedure?
- Not Applicable

Have standard biosecurity and institutional safety procedures been adhered to?
- Not Applicable

Q5:

This study presented two models (GAM and LSTM) for bias correction (caused by environmental parameters change) on measurements from two instruments.

The results in the manuscript demonstrated that both models could significantly reduce the biases.

However, there are issues that need to be addressed, which are also listed in Q2:
1. The hyper parameters for the LSTM were determined in a quasi-empirical approach. The result may not be optimal.

We have adopted the GridSearch approach that does a systematic permutation of hyperparameters. Please see our response to Point #3, below.

2. The review could not identify the details of how to separate the data set into training and testing sets.

We set the length of the test dataset as twice the number of timesteps (t), to ensure that the test data set can capture periodic data contained within the timestep interval. The remainder of the data set is used as training data. This is describbed at the end of section 3.2 near lines 474-477.

3. K-fold cross-validation should be adopted when training and testing the models to have more comprehensive results.

We have implemented k = 5 fold cross-validation using the GridSearchCV wrapper from scikit learn. The GridSearch was used to probe all permutations of the epochs, batch_size, and dropout hyperparameters. The dropout likelihood has been added at this stage to better mitigate overfitting. The technique of randomly removing network nodes from the LSTM solution has the effect of creating a more general grid and regularizing the network weights so that none dominate.

We would like to point out that the ultimate demonstration of the bias correction method is in computation of the RMSD against the independent instruments (SBE43 and infrared spectrometers). Because these data sets are independent of the training data, there is no merit to randomly resampling the data used to validate the fit.

4. In Figure 9, the LSTM began overfitting when epoch number was above 10; efforts such as reducing the network complexity should be taken in order to prevent overfitting instead of simply stop at epoch 5.

Thank you for the comment. After implementing the gridsearch, it was determined that Epochs = 20 produced the optimal RMSD fit. We have updated Figure 8 to reflect these changes. By the same token, we have eliminated Figure 10 which did not make a significant contribution to the analysis of the LSTM solution.

5. The review suggested a comparison base line such as simple multi-variable regression model should be included.

We take the reviewer's point, that comparison of the methods with a simple base line or benchmark might be instructive to some readers. However, there are many similar tutorials of both methods that are available in the literature and in the public domain. Our experience with both fitting approaches is that the GAM and RNN will be highly skilled at fitting to a simple multivariate function, so there is little to be gained in discriminating between the two methods.

6. Lacking a comprehensive literature review.

We acknowledge the point about a thorough literature review, and we have included what we feel are the most defining texts in the area of guided machine learning as well as references to the practical guides that we used for implementation. In the field of instrument bias correction, we did not find many relevant or supporting texts. We feel that an exhaustive review of machine learning and multivariate theory would unnecessarily weigh down this manuscript, which is focused on presenting an application.

Line 687, “Panels c e and f show the…” should be “Figure 3. Panels d e and f show the…”

Fixed, thank you.

Line 214, “which took place ca. every other day,”, is “ca.” a typo?

Ca. is short for circa and means approximately, but maybe this is not a common usage, so I have changed to approximately, to clarify the meaning.

Line 718, typos at the end.

Fixed, thank you.

Line 424, “Sample” should be “sample”.

Fixed, thank you.

Line 461, typo “t” with under line.

Fixed, thank you.

Line 509, duplicate “a”.

Fixed, thank you.

Line 548, typo “µatm”.

I don’t see the typo, I have reviewed this line. The use of “µatm” is correct in the version of the manuscript that I am looking at and in the pdf.

Line 549, missing unit after “35.2”

Fixed, thank you.