Review of Wang et al., Global ocean dimethyl sulfide climatology estimated from observations and an artificial neural network.

This manuscript describes a novel methodology for deriving a global ocean dimethyl sulfide (DMS) climatology, using an artificial neural network (ANN). The authors demonstrate that the ANN is able to explain a greater fraction of variance in the raw available observations of surface ocean DMS concentrations, as compared with a multiple linear regression approach. They also contrast this approach with previous work that used spatial and temporal gap-filling to estimate DMS concentrations, including in data-sparse regions. Instead, the approach presented here derives relationships between observed environmental parameters and observed oceanic DMS DMS concentrations (using the multiple regression or ANN), and uses these to predict/extrapolate DMS concentrations globally.

The paper is clearly written, the methods are straightforward and appropriate, and it represents a valuable contribution to work on understanding and representing the present-day climatological distribution of DMS concentrations in the surface ocean. Improved climatologies of DMS would be useful for Earth System models, especially if they can offer more insights into how the DMS production would change under past/future climate states. It's unclear (to me, at least) whether a machine learning approach will be able to offer such physical insights. Nevertheless, such approaches can offer a better estimate of the present-day state, and this is useful in itself for Earth System modeling. The uncertainty in ocean DMS climatologies is still quite large, despite advances during the past decade, and new advances in statistical approaches that can reduce errors in these datasets are welcome.

Thank you for your positive comments.

I have only a few minor comments, as follows:

I agree with the comments of the two previous reviewers that the arguments made against data binning are weak. The authors imply that it is an inherently inferior approach, but, this is not necessarily true a prior. There can be good arguments in favor of data binning before analysis, e.g., to harmonize the temporal and spatial scales of multiple datasets before analyzing the relationships between them. When in situ DMS measurements (essentially instantaneous) are being predicted via monthly mean values of chl-a, MLD, etc., it is not at all obvious that it is appropriate to perform the analysis without first binning the data. This point should be treated with more nuance, taking into account the details of the datasets and the processes involved.

This point was also raised by the other two reviewers. We therefore dealt with it very carefully, and added the following arguments.

The PMEL database expanded dramatically. Now there are a total of 86,785 valid DMS measurements (concentration greater than 0.1 nM and less than 100 nM according to your instructions), that is 71% larger than the number of data we initially used (51,161). For the expanded data set, ~93% of DMS are accompanied with in-situ SST measurements, ~81% are accompanied with in-situ salinity measurements. More importantly, each data point has their unique location and sampling time signatures. As shown in the following figure, sampling time (date) and location information is a strong DMS predictor, which together can decrease DMS root mean square error to 0.64 (on natural logarithm scale). Adding other climatological predictors can further improve the model performance.

The NAAMES dataset has 6,786 valid data points, which are ~7% of the total data points (93,571 = 86785+6786). All data are accompanied with in-situ Chl a, SST, and SAL measurements. For parameters without in-situ measurements, high resolution data are used to match DMS measurements, 0.0417° ×0.0417° for PAR, 0.5° ×0.5° for MLD, and 1° ×1° for NO3, which ensures most of DMS have a set of unique predictors. As shown in Table 1, merging NAAMES data with PMEL data does not significantly change the statistic.

Moreover, binning the data will reduce data variance, which has been demonstrated by Derevianko et al. (2009). The objective of this study is to train an ANN with as many data as possible, so that the model is generalized. It not only can apply to coarse resolution predictor fields, but also can apply to very fine resolution field, for example, we have applied the network to fine resolution NAAMES fields for comparison with in-situ DMS measurements (Bell et al., in prep.).

Lastly, binning data will also result in loss of information. A great amount of information is associated with sampling time and date as shown in the following figure (Fig. 2a in MS). By binning the data into monthly 1° × 1° grid, the valid DMS data points will decrease significantly from 82,996 to 9,018; sampling date feature (365) will be average to 12 months, and coordination combinations will be averaged from 87,332×87,332 to 180°×360°, which represents great information reduction. For ANN models, less data points usually lead to overfitting Fig. 2b.

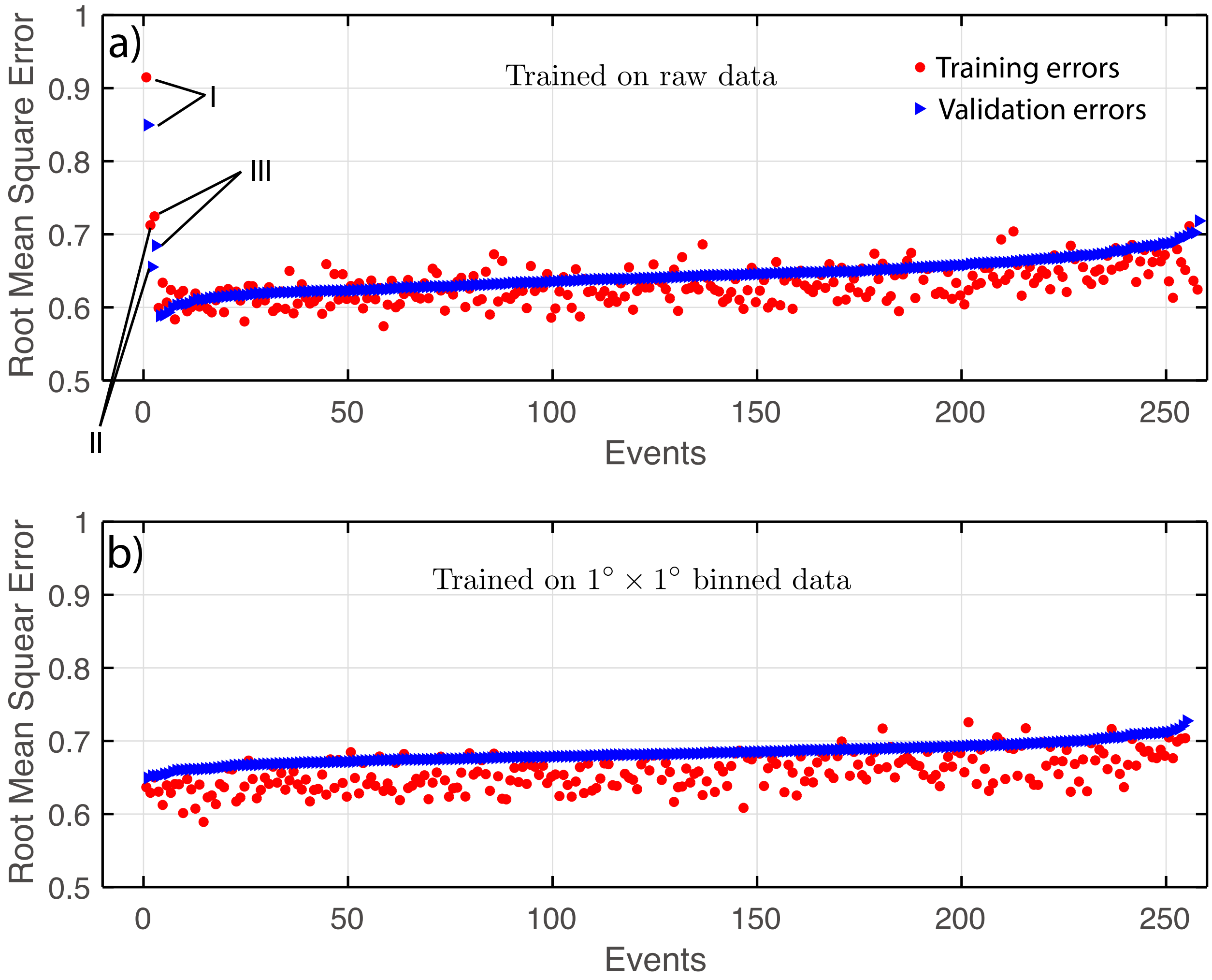

Fig. 2 Parameter sensitivity tests on raw and binned data. (a) Root mean square error on logarithmic scale for the model trained using raw data; (b) Root mean square error on logarithmic scale for the model trained using binned data. The time and location parameters are tested separately without combining with environmental parameters as shown in the upper panel, (I) with only location parameters; (II) with location and day of year parameters; and (III) with location, day of year, and time of day parameters. The model with three location parameters (I) has a root mean square error on natural logarithmic scale of ~0.83, which decreases to ~0.65 by adding sampling day of year parameters (II), however, increases to ~0.67 by adding sampling time parameters (III). We, therefore, do not include sampling time parameters in the following tests. We tested every possible combination of the eight parameters (PAR, MLD, SST, SAL, Chl a, DIP, DIN, and SiO), which in total are 255 tests.

p. 5, l. 128-130: I was glad to see that the authors have considered the issue of potential overfishing, but they don't explain how they determined that the setup they used for the ANN is not overfitting (i.e., what methods or criteria were used to determine this). It's common to use multiple rounds of cross-validation (such as k-fold crossvalidation or related methods) in order to determine whether a statistical model may be overfitting and to assess the uncertainty in the fit. If I am understanding the description of the method correctly, it seems that while the authors divided the data into training and validation subsets, they did so only once. In this case, the results of the ANN will be sensitive to the specific subset of data that was used for training it. It should be explained how the training/validation subsets were selected, and also whether a multiround cross-validation method was employed (and if not, why not). Or, if appropriate, the authors could simply carry out a more thorough cross-validation and update the manuscript, since I expect this should not require much effort.

Good point.
There are two general guidelines when one separates the data to training and validating sets, representation and generalization. That is to say that your training data has to be representative, and your model has to have the ability to generalize. The online DMS data are organized by contributor ID, while when you do cross validations, the data are drawn section by section as the following figure shows. One section of data may be from a specific contributor who collected data from a specific region, therefore, the data may not be representative, which results in an over-trained or a less-trained model (we have uploaded the cross-validation model to github ((https://github.com/weileiw/ANN-DMS-code), so interested readers can play with it.). To make the selection more representative, a common practice is to shuffle the data, and then randomly draw a fraction from the shuffled data. For DMS data (or maybe other oceanography data too), data collected from the same cruise are highly intercorrelated, so that shuffling and randomly splitting will "leak" information to the model and cause an overfitting (we have tested shuffling and random drawing method, it indeed leads to overfitting. (Code is also available at Github directory.)) .

Another purpose of doing cross-validation is to allow your model to see as many data as possible. This is useful when you do not have enough data to train your network. To achieve a similar effect, we first manually adjust the hyper-parameters (dropout ratio, hidden layers, number of nodes etc., they are key parameters to determine the model performance) using manually-divided training, internal testing, and external validation data. After we get a satisfactory combination of those hyper-parameters, we fix them and fine tune the network using all available data (because the data are intercorrelated, shuffle and randomly split training and testing does the work.).

Lastly, in the parameter selection experiments, we examined a total of 255 models (every combination of eight environmental parameters). We then ranked the model according to root mean squared error (RMSE) on validation data as shown in Fig. 2a. Compared to RMSEs on the training data, there are no apparent overfittings for the top 10 models. The models with larger RMSEs generally overfit the training data. Meanwhile, overfitting occurs with almost every model when binned data are used (Fig.2b).

Accordingly, we have added more explanations in the revised MS as follows (l.154 – l.162):

“The data was split into sets manually rather than automatically. The online DMS data are organized by contributor ID, and automatic splitting draws a continuous portion from the data. The data portion may come from a specific contributor who collected data from a specific region and it may therefore not be representative. This would result in an over-trained or a under-trained model. To make the selected data more representative, a common practice is to shuffle the data, and then randomly draw a fraction from the shuffled data. For DMS, data collected from the same cruise are highly intercorrelated, so that shuffling and randomly splitting "leaks" information to the model and causes overfitting. We manually adjust the hyper-parameters (dropout ratio, hidden layers, number of nodes etc.) using the data that has been manually-divided into training, internal testing, and external validation subsets. After obtaining a satisfactory combination of those hyper-parameters (as discussed below), we fix them and fine tune the network using all available data.”

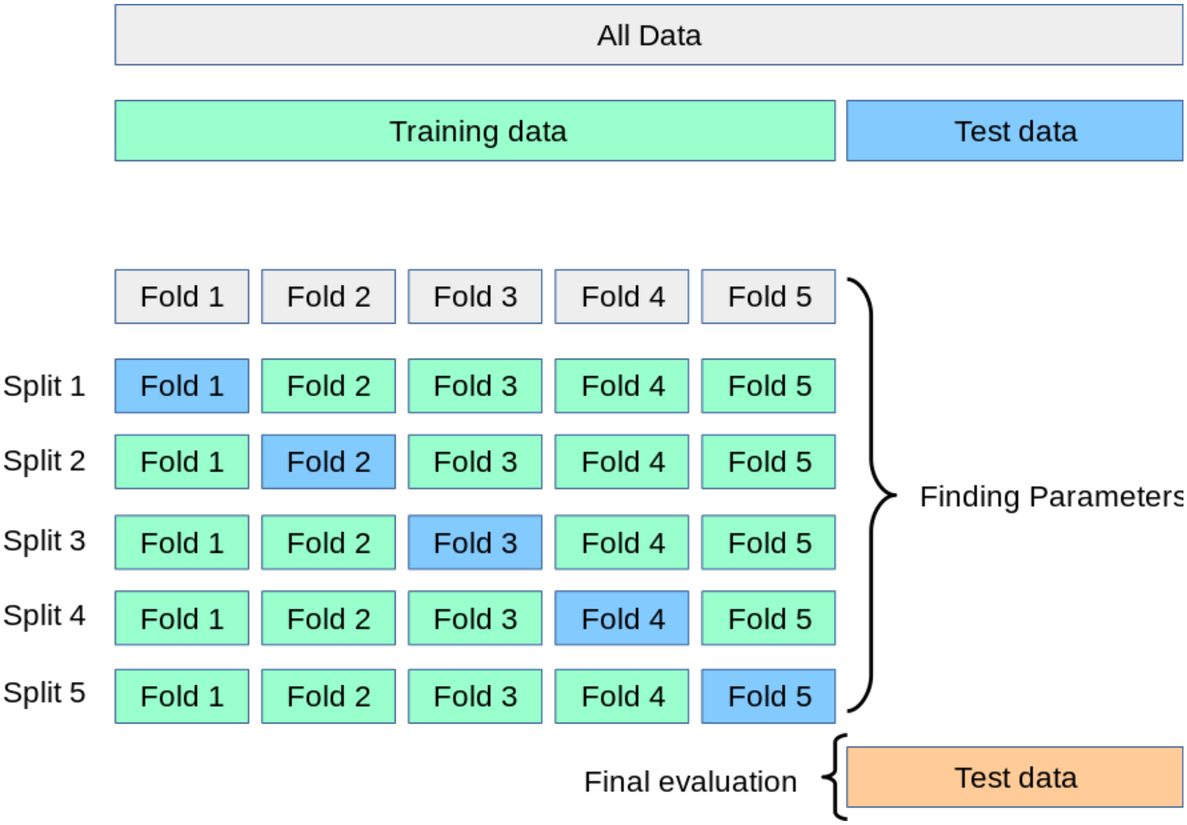
All Data
Training data
Test data
Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Split 1 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Split 2 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Split 3 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Split 4 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Split 5 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Finding Parameters
Final evaluation
Test data

Figure cited from https://scikit-learn.org/stable/modules/cross_validation.html.

p. 5, l. 133-134: It was not obvious to me what the “random states” refer to – is this a random seed controlling initial parameter values?

This is a good point and following is the explanation.
In the ANN, there are at least two places using random states, 1) it uses random state to decide the Dropout nodes, 2) it uses random state to separate internal testing data from training data. The random states do not control initial parameter values, but different random states produce slightly different results. To make our model results reproduceable, we fixed the random state at 64 in the revised model. The uncertainly analyses are now based on different parameter combinations.

p. 8, l. 220: here, it is stated that ANN is able to “capture more of the variance” than “previous extrapolations (Kettle et al., 1999; Lana et al., 2011)”. This is a key claim of the paper in terms of the claimed improvement over previous methods, and I can believe this is probably true, but I think the claim ought to be supported by a quantitative value – i.e., the percentage of variance captured by the two previous climatologies – so that readers can compare and see the improvement in this metric. Perhaps these values are in the manuscript somewhere and I overlooked them – in that case I think they should be featured somewhere that is easier to find (e.g., in the abstract or in a table).

Good point.
However, it is hard to do an apple-to-apple comparison. Because we are comparing to raw data, whereas, Kettle et al., 1999 and Lana et al., 2011 interpolated the data, it is hard to extract the raw data information from the climatological map. We thus changed the wording, and weakened the comparison as follow,

“the ability of the ANN to build a nonlinear relationship between DMS and environmental predictors allows it to capture much of the variance”

We also added more comparison to previous model results as shown in Fig. 3. and Fig. 4 in the text, also attached below.

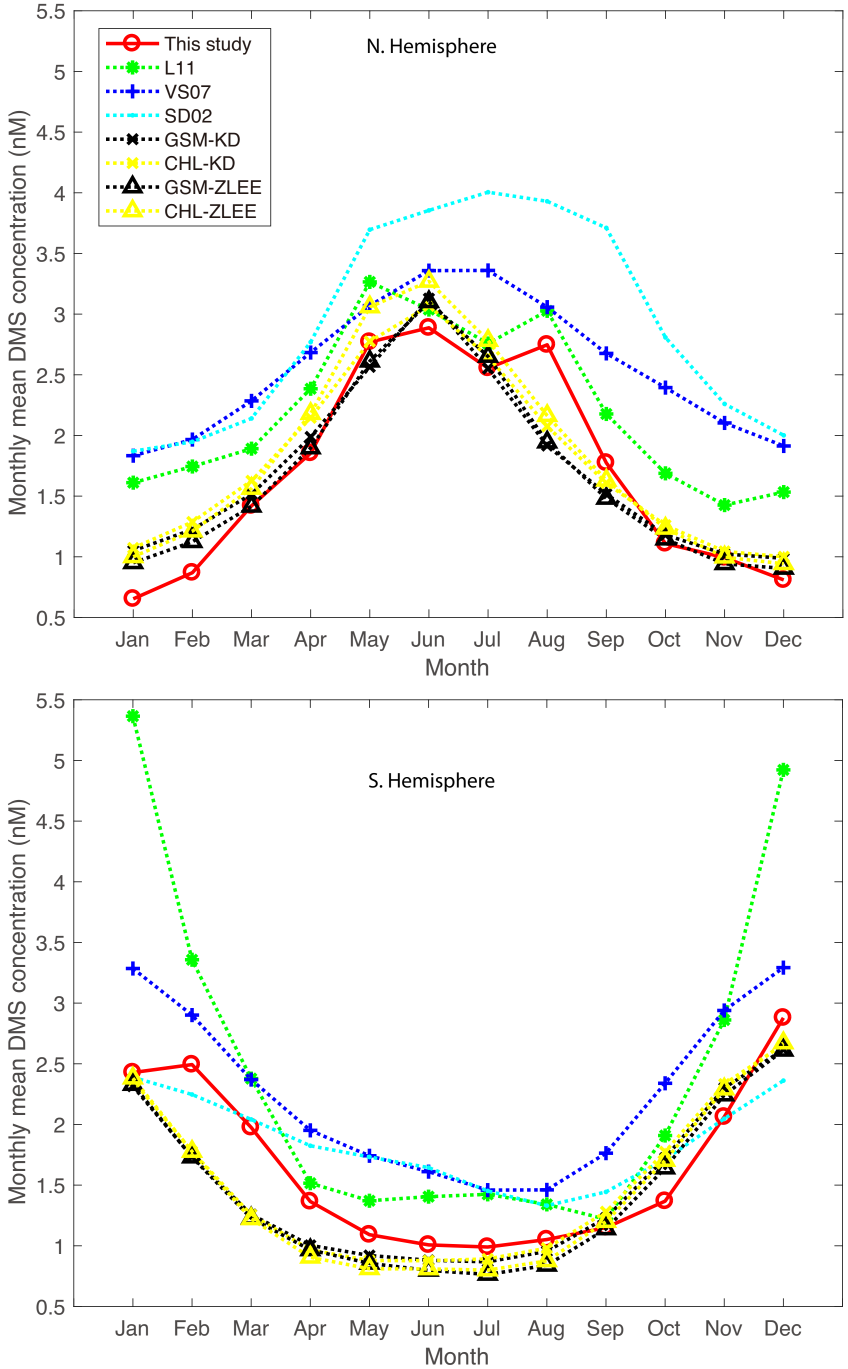

Figure 3. Comparisons of monthly mean DMS concentrations between this study and previous studies (Simó and Dachs, 2002; Vallina and Simó, 2007; Lana et al., 2011; Galí et al., 2018).

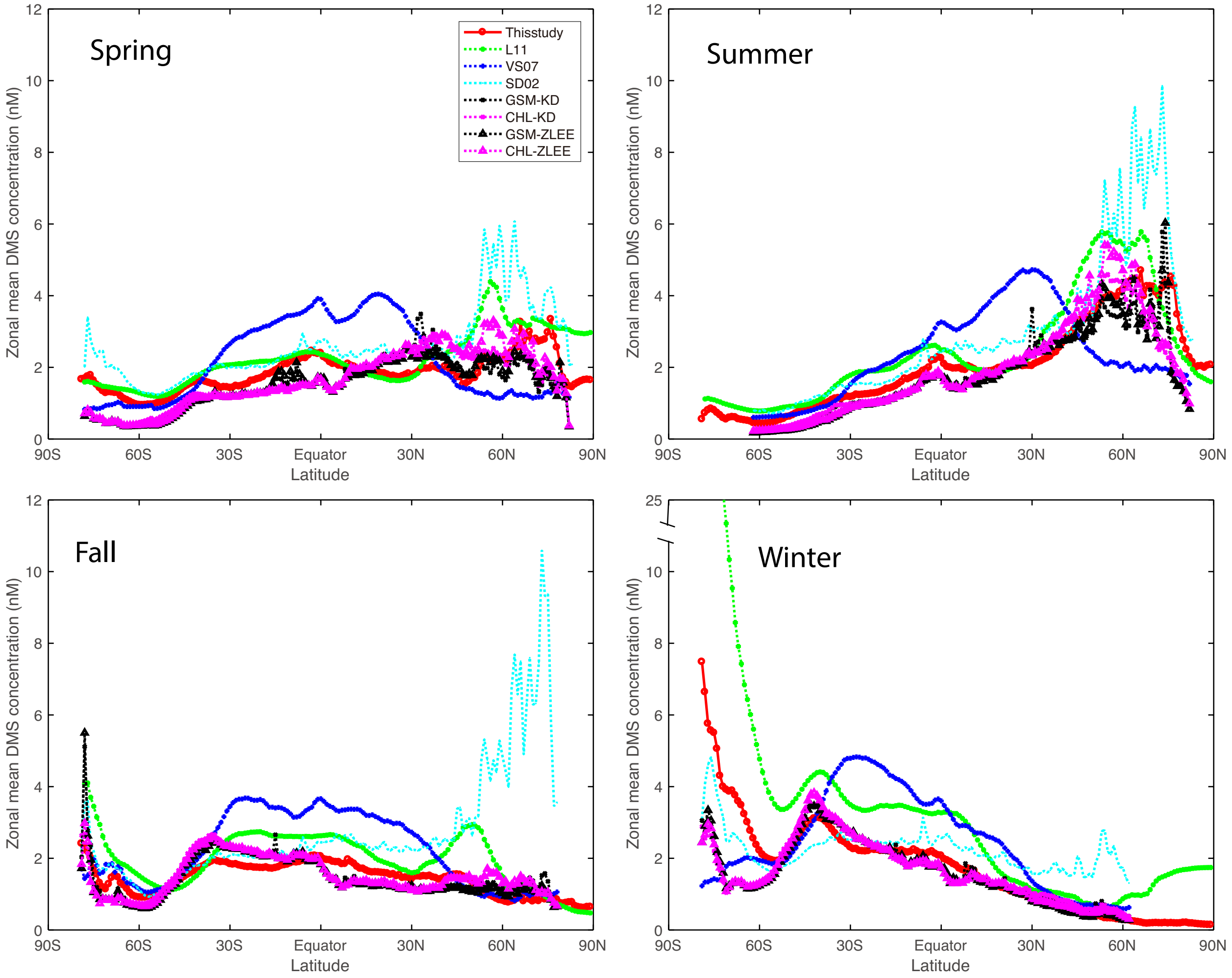

Figure 4. Comparisons of zonally mean DMS concentrations between this study and previous studies (Simó and Dachs, 2002; Vallina and Simó, 2007; Lana et al., 2011; Galí et al., 2018).

p. 11: I tested the links for the code and data availability; the data doi link at zenodo works, but the github link does not seem to be available.

The code is previously in a private repository, and now is public (https://github.com/weileiw/ANN-DMS-code). We have also uploaded the corresponding data used to train the model in the following directory: https://zenodo.org/record/3833233#.XsM4cBP0nV4

I also noticed a couple of typos:
p. 2, l. 40: "result" -> "results" or "result[s]"
Corrected, Thank you.
p. 5, l. 31: "deduction" -> "reduction"
Corrected, Thank you.

p. 5, l. 133: “assemble” -> “ensemble” (?)
Corrected, Thank you.
p. 7, l. 189: “wasters” -> “waters”
Corrected, Thank you.