Dear Dr. Murat Aydin,

Thank you very much for reviewing our revised manuscript. Your insightful comments are very helpful in further improving the quality and completeness of our work. Please find our replies below in blue. The proposed changes in manuscript are shown in green. Specified line numbers before and inside the bracket refer to those in revised manuscript with and without track of changes, respectively.

Best regards,

Ying Chen on behalf of all authors

I acknowledge the extensive nature of revisions to the paper. While many of the revisions resulted in improvements, some feel like a step backward. Given that this the second go around, I will not dwell on minor issues. I do recognize the value of the ANN data set in terms of high resolution in both temporal and spatial scales. My main concerns are related to the monthly and annual fluxes, specifically the fidelity of these estimates to reality as defined by the available observations.

There is only one way to test the accuracy of the ANN data product: it has to be compared with the DMS obs that underlie the training of the machine learning process. In my first review, I suspected that the linear regressions between the DMS obs and the ANN estimates yielded slopes significantly different than 1 and suggested the residuals might be correlated with the observations. I further added that the statistical metrics they relied on were insufficient to adequately evaluate the accuracy of the data product. The additional analyses the authors conducted based on the review confirm my suspicions were correct. While I appreciate the effort that went into the revisions, I do have misgivings about a major aspect revision they implemented and suggest further revisions.

The weighing scheme implemented to increase the influence of low and high concentrations on the results is a data analysis gimmick aimed at improving the linear regressions with respect to the deficiencies I outlined in the first review. I do not believe it is appropriate to manipulate the distribution of the training data in this manner unless they are real life reasons (related to the real world ocean and how it has been sampled) why lower and higher DMS concentrations are underrepresented in the observational data sets.

The manuscript offers no such justification. As such, they would be better of presenting the original ANN results as the main data product and offer the weighing-based results as supplementary analysis. When referring to this supplementary analysis, you should discuss in the main body of the manuscript why it was conducted. In my view, the implemented weighing scheme does not make enough of a difference in the end and I remain unconvinced that the problematic aspects of the linear regressions are caused by extreme concentrations that constitute a small fraction of the data set. There appears to be a systematic issue for reasons that remain unclear to this reviewer.

Thanks for your comments on this issue. The weighted resampling strategy or over-sampling of the minority class is a widely used approach in machine learning to deal with data imbalance and improve the model performance and generalization (Haibo et al., 2008; Yu and Zhou, 2021; Chawla et al., 2002). During the ANN training process, the model tends to focus on optimizing the majority class of data, such as samples with moderate DMS concentrations in this study, and may overlook data patterns within the minority class. Increasing the proportion of the minority class in the training process can let the model learn more information from these samples. However, given the limited improvements in this study, we acknowledge that there are other important issues contributing to the systematic bias. Potential reasons include (1) a mismatch in the spatial and temporal scales between the input and target, (2) uncertainties associated with the input data and DMS measurements, (3) limited capability of machine learning model to fully capture the complex input-output relationships, and (4) the effects of other environmental factors not incorporated in this study. The details are discussed in Section 4 (***Uncertainties and limitations***).

Here we have reverted to using the original ANN model construction to generate the DMS data product. However, some updates made in the second version, including the update of input data sources and the inclusion of more DMS observations, are retained. The results after implementing the weighted resampling scheme have been moved to supporting information and regarded as an approach to test whether the systematic bias is attributable to data imbalance. The details of how the weighted resampling scheme was conducted have been moved to Appendix. The simulated DMS distributions does not show significant differences compared with the second version. All figures and values in the manuscript have been updated.

Lines 273-288 (246-261): However, it is noteworthy that our model tends to underestimate extremely high DMS concentrations and overestimate extremely low concentrations. Overall, the linear regressions between ANN-predicted and observed DMS concentrations yield slopes significantly lower than unity across all regions (Fig. 3c and 4), and there are significantly positive correlations between prediction residuals (observation – prediction) and observed $\log_{10}$(DMS) (Fig. S5 and S6). From a data perspective,

this may be partly due to the insufficient number of samples with extreme DMS concentrations (known as underrepresentation), making it difficult to adequately capture the relevant information during training process. To test this point, we adopted a weighted resampling strategy to bolster the number of samples in the minority class before training, which has been widely used in machine learning to deal with the data imbalance issue (Haibo et al., 2008; Yu and Zhou, 2021; Chawla et al., 2002). The basic idea is to set a higher probability of being sampled for the minority class with extreme DMS concentrations, and the details are illustrated in Fig. S7 and explained in Appendix B. The results indicate that the weighted resampling scheme cannot fully alleviate the model bias. Although it does elevate the overall prediction-versus-observation slopes from ~0.59 to ~0.63, this improvement is marginal (Fig. S8 and S9). In several regions like Westerlies_S and Trades biomes, the slopes are even lower than original values. Furthermore, the data become more scattered after implementing the weighted resampling, resulting in increased RMSE and decreased R^2. Therefore, there are other potential issues causing the model bias, which are discussed in Section 4. The original model, trained without weighted resampling, was adopted for subsequent analysis and the construction of the gridded DMS dataset.

Lines 594-608 (538-552): Although our ANN ensemble model and derived DMS dataset demonstrate certain advantages compared to previous studies, as discussed in Section 3.3, there persist notable uncertainties and limitations, which result in the ~35% uncaptured variance (Fig. 3a) and non-negligible simulation biases, e.g., underestimation of extremely high DMS concentrations and overestimation of low DMS concentrations. Firstly, there is a mismatch in the spatial and temporal scales between the input and target. The target, sea surface DMS concentrations, are obtained from in-situ measurements taken at specific locations and time points. In contrast, the input data are primarily from gridded datasets where each pixel represents an average over a defined spatial and temporal range. This is particularly significant for the ECCO variables, which have the largest spatial grid size of 110 km. Consequently, extreme values at specific locations cannot be accurately captured by the regional averages, resulting in dampened variations among the samples. Secondly, the input data from different sources and the observed sea surface DMS concentrations inherently possess certain uncertainties, which can introduce noises into the ANN learning process. Thirdly, the ANN itself may not be powerful enough to fully capture the complex input-output relationships across different oceanic regions, especially when the samples are scarce under specific environmental conditions. Finally, beyond the 9 variables incorporated in this study, other environmental parameters such as pH (Six et al., 2013; Hopkins et al., 2010) and trace metal elements (Li et al., 2021) can also influence DMS concentration. Not incorporating these factors may introduce additional biases.

Lines 701-722 (621-641): **Appendix B: The weighted resampling strategy**

Apart from the data imbalance between coastal and non-coastal regions, there exists an imbalance across different DMS concentration ranges. The majority of DMS concentrations (78.6%) fall within the range of 0.8 to 10 nM ($\log_{10}$(DMS) between -0.1 to 1). Samples with DMS concentrations exceeding 15 nM or falling below 0.3 nM only represent 6.9% of the entire sample set. A weighted resampling strategy was applied to mitigate this imbalance (Fig. S7). We randomly sampled 50,000 samples with replacement from the original sample set. The probability of each sample being selected is proportional to the weighting factor shown as the red dash line in Fig. S7b, which is dependent on its DMS concentration. First, the probability distribution of initial $\log_{10}$(DMS) values was fitting with a gamma distribution, which is given below and displayed as the blue line in Fig. S7b:

$$f(x) = \frac{1}{\Gamma(k)\theta^k}(x+4)^{k-1}e^{-(x+4)/\theta} \quad \text{(A1)}$$

Here k and θ represent the shape parameter and scale parameter, in this case, 100.7 and 0.044, respectively. x is the $\log_{10}$(DMS) value. Since gamma distribution only takes positive values, we added 4 to the original x as the dependent variable for distribution fitting. We then obtained a new gamma distribution function with the same mode but lower shape parameter, in which $k = 40$ and $\theta = 0.112$. The reciprocal of the new gamma distribution function was taken as the weighting factor. As a result, samples exhibiting high or low DMS concentration values are more likely to be selected, whereas those with intermediate concentrations are less likely to be selected. We also controlled the F_{coastal} value of the resampled data equal to 9.7%. The data distribution of DMS concentrations after the resampling process is shown in Fig. S7c. The fraction of samples with DMS concentrations above 15 nM or below 0.3 nM is elevated to 15.0%. The 50,000 samples were then randomly split to a training set (80%) and a validation set (20%). Since there are duplicate samples in the resampled dataset, the random data split was conducted based on the original sample ID before resampling to ensure that there was no sample overlap between the training and validation sets.

Further, I do not like the fact that the comparison of the training data versus the observations are not shown in the main manuscript anymore. If the number of figures in the manuscript is a problem, I suggest moving the residual figures to the supplement and showing the main comparison figures with respect to both the training and test data in the main body. The slope values should be displayed in all sets of figures. Most readers may not readily infer the implications of trending residuals and the manuscript does not offer a detailed enough discussion.

Thank you for your suggestions. We have moved the comparison for training set back to the main text and placed the residual plots in the supporting information. The slopes of linear regression between predictions and observations are also added in the figures. The discussions for the potential reasons of slopes lower than unity are provided in Section 4, as mentioned in the response to the above comment.

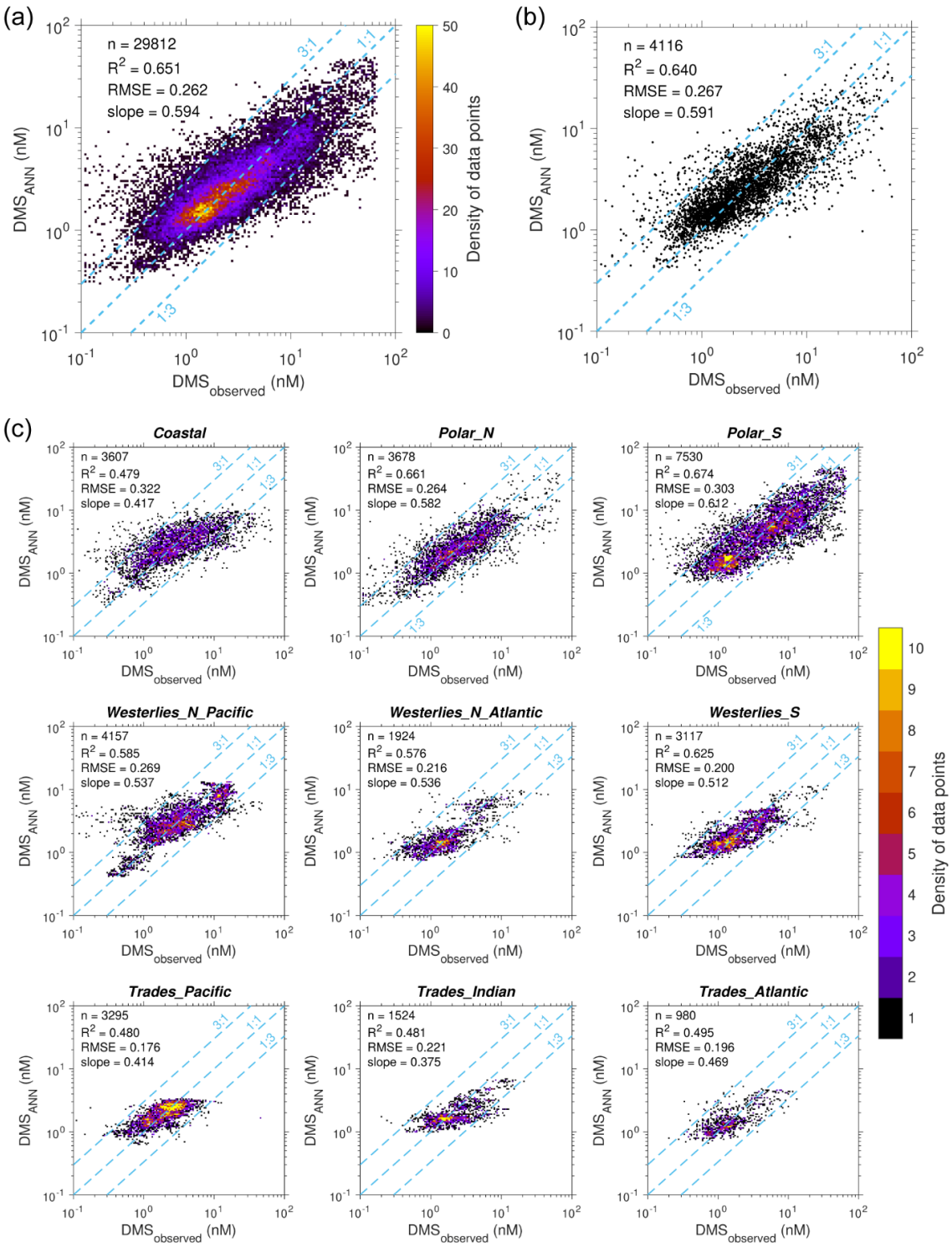

Figure 3. Comparisons between ANN-simulated and observed DMS concentrations. (a) Scatter density for simulated versus observed DMS concentrations of the samples used in ANN training. (b) Comparison between the simulated versus observed DMS concentrations of testing set. (c) Comparison between the simulated versus observed DMS concentrations of the samples used in ANN training across 9 regions. The number of data points (n), $\log_{10}$ space R^2, root mean square error (RMSE), and linear regression slope are also displayed.

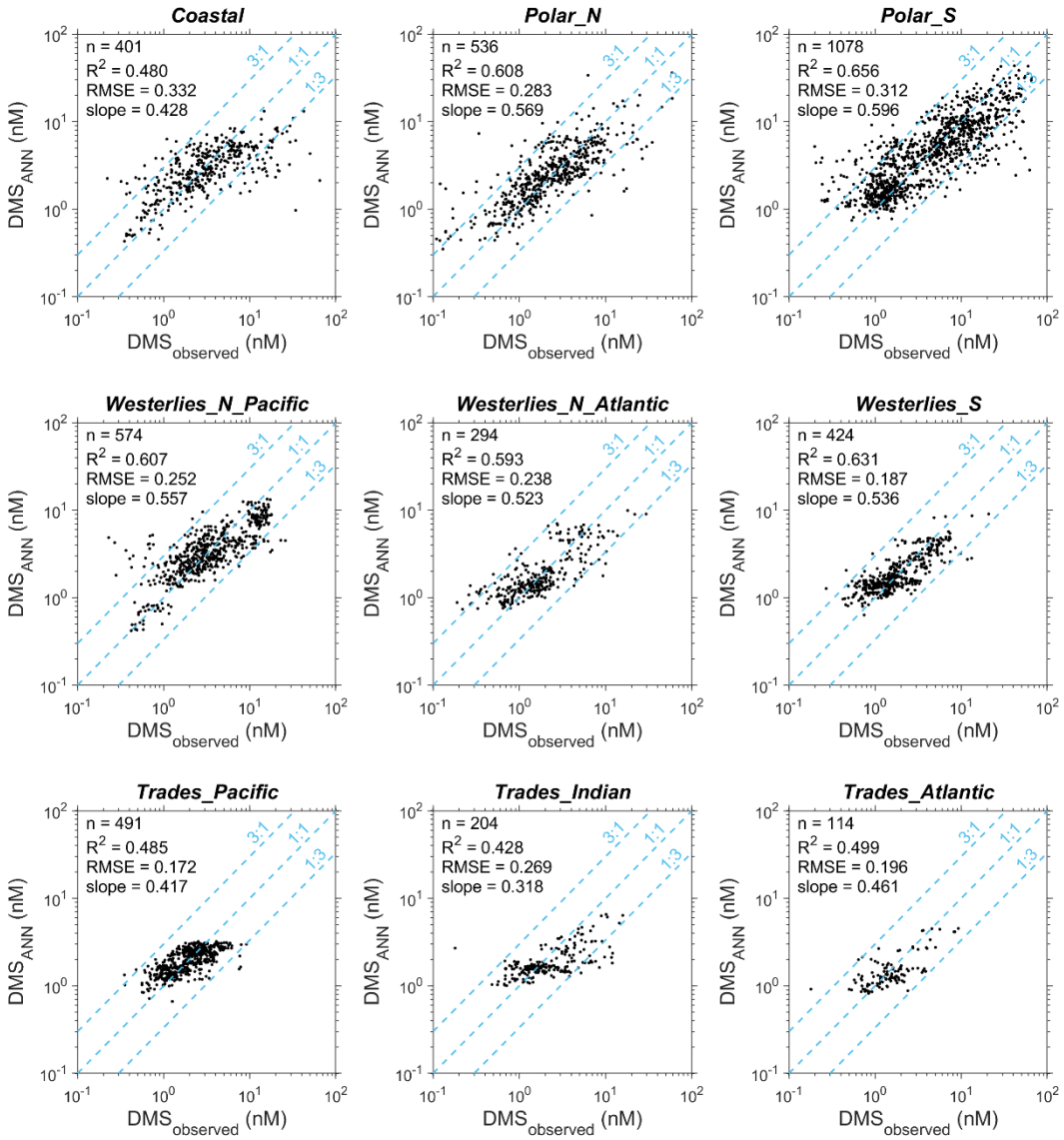

Figure 4. Comparisons between the simulated versus observed DMS concentrations of the testing set across 9 regions.

A welcome revision to the manuscript is the inclusion of regional mean and normalized mean bias estimates presented in Table 2. However, this is the bare minimum necessary since the positive and negative biases that occur at high and low ends of the concentrations tend to cancel out during the averaging, therefore hindering insight into the biases at grid scale let alone how these biases impact the regional and global fluxes. I'm willing to accept these outstanding issues as subjects of future work as long as they are pointed out in the paper.

Thanks for your comments. We acknowledge that this is a critical issue needs to be addressed in the future. We have pointed out that the negative biases at high end of the concentrations will be partially cancelled out by the positive biases at low end during the averaging and the bias at a specific grid could be much larger. In Section 4, we have also proposed several measures that can be taken in the future to mitigate this bias.

Lines 306-308 (268-269): On the other hand, the negative biases at high end of the concentrations are partially cancelled out by the positive biases at low end during the averaging over the entire region. The bias at a specific grid could be much larger.

Lines 609-622 (553-565): The overall bias for $\log_{10}$DMS is at a similar level between high- and low-concentration ends, but the DMS concentration on a linear scale is more underestimated in the high-concentration regime than it is overestimated in the low-concentration regime. As a result, our simulation results may tend to underestimate the annual average DMS concentration and flux. To mitigate this critical bias and reduce model uncertainty, high-quality input datasets with finer spatial resolution are needed in the future. The high-time resolution nature of the resulted daily DMS data product would be more valuable if accompanied by higher spatial resolution. Expanding the data volume is also crucial for improving model performance. Although the current DMS observational data covers all major oceanic basins, certain regions such as the Trades_Pacific remain underrepresented. Advances in online measurement technologies offer promising avenues for acquiring more extensive and convenient observational data (Hulswar et al., 2022). Additionally, incorporating more input features to the model would be beneficial. This necessitates a comprehensive understanding of the spatiotemporal distributions of those input features, and further field measurements are important to this end. Moreover, integrating DMS biogeochemical mechanisms with machine learning technique, i.e., a hybrid model coupling physical processes with data-driven approach, may further improve prediction accuracy, generalization, and interpretability (Reichstein et al., 2019).

References:

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P.: SMOTE: synthetic minority over-sampling technique, Journal of artificial intelligence research, 16, 321-357, 2002.

Haibo, H., Yang, B., Garcia, E. A., and Shutao, L.: ADASYN: Adaptive synthetic sampling approach for imbalanced learning, 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008, 1322-1328.

Yu, L., and Zhou, N.: Survey of imbalanced data methodologies, arXiv preprint arXiv:2104.02240, 2021.