Answers to Referee 1

We would like to express our gratitude to the reviewer for their insightful comments on the manuscript. According to the comments, the manuscript has been revised. Please find our replies in the pdf file. In the responses, the reviewer's comments are in black text, and our responses are in blue and *the main text modifications to the revised manuscript are in italics*.

Mansour et al. used the machine-learning model to predict the biogenic methanesulfonic acid (MSA) and sulfate (SO4) concentrations covering the North Atlantic Ocean. Overall, the study is very interesting and falls into the scope of ESSD. However, the manuscript still suffers from some major weaknesses. I recommend the manuscript for publication on ESSD after the following comments have been well addressed.

1. The novelty of this dataset in this study should be well clarified in the introduction.

The introduction has been updated as follows:

"In this study, we present the first high-resolution and long-term daily gridded time series of freshly formed In-situ Produced Biogenic Methanesulfonic Acid and Sulfate (IPB-MSA&SO4) concentrations over the NA ocean at 0.25° × 0.25° spatial resolution. The data covers 25 years from 1998 to 2022 with the possibility of future updating year by year. The dataset is a unique and novel product that in fact extends the space and time representativeness of atmospheric in-situ observations of marine aerosol chemical properties over the North Atlantic Ocean, by exploiting the potential of machine learning. The dataset indeed represents the sea-level concentrations of MSA and SO4, in each grid point of the domain, resulting from the interplay between precursor emissions and local atmospheric conditions."

In addition, the following sentence has been added as a complement to the 2nd paragraph:

"Recently, multilinear regression was utilized to simulate monthly MSA over the eastern China seas at a spatial resolution of 1° × 1° (Zhou et al., 2023), concluding that MSA spatial/seasonal patterns exhibit significant variability, which is primarily governed by surface phytoplankton biomass and the atmospheric boundary layer height."

2. Why do not you use the physical model (e.g., MITgcm, ROMS) output as the input for the machine-learning model? As shown in figure 1, the middle region of Atlantic lacks of measurement, and this region might show large uncertainties based on machine-learning models alone.

Here it is necessary to clarify that the only data missing in the central part of the domain (Fig. 1) are the atmospheric observations of MSA and nss-$SO_4^=$ concentrations. These cannot be accessed from the output of the physical models suggested by the reviewer. Furthermore, the aim of this work is to provide an alternative to uncertain model outputs of MSA and $SO_4^=$ concentrations, by using an original approach to extend the time and space representativeness of in-situ measurements. All the other data used to generate the IPB-MSA&SO_4 dataset come from satellite observations or previously generated datasets which cover the whole domain. For instance, one of the main predictors used for calculating MSA and $SO_4^=$ atmospheric concentrations, is the sea-to-air DMS flux (F_{DMS}), serving as the main tracer of marine biogenic sulfur aerosol concentrations in the atmosphere. The F_{DMS} data product has been parametrized from seawater DMS concentrations and wind speed (Mansour et al., 2023). Ultimately, F_{DMS} is a daily (0.25° × 0.25°)– Level 4 product, covering the whole North Atlantic domain, which means that the considered domain has no missing data points. The DMS has been reconstructed by merging high-resolution satellite data (chlorophyll-a concentration, sea surface temperature, and photosynthetically active radiation), Ocean Physics Reanalysis (oceanic mixed layer depth and seawater salinity) and Ocean Biogeochemistry Hindcast (sea surface concentrations of nitrate and silicate).

As the reviewer, we were also worried that the scarcity of the MSA and $SO_4^=$ observational data in the central part of the domain may have resulted in biased (less constrained) predictions in such part of the domain. As a confirmation of the validity of the implemented GPR model, we evaluated the possibility of reconstructing daily variations of nss-$SO_4^=$ during NAAMES campaigns (in the westernmost part of the study area) in the worst-case scenario of training the GPR model only using the MHD data (i.e., measurements collected at the Easternmost side of the domain). This exercise provides an idea of the reliability of the ML approach to model MSA and $SO_4^=$ concentrations in regions poorly constrained by in-situ observations. The results are shown in the following figure displaying that GPR, also in this worst-case scenario deployment, can explain 55% of the daily observed nss-$SO_4^=$ variance. It is also worth considering that the dataset used for this test is limited (*n=57, as the days of observations available from NAAMES*) and does not cover a full seasonal cycle , which makes it harder for the model to perform a good prediction as the seasonality is the main

driver of the variability in biogenic aerosol emissions in the studied domain. Anyhow, we consider the performance of GPR in this worst-case scenario test as more than acceptable and believe consequently that the IPB-MSA&SO4 dataset may be considered reliable also in the central part of the NA, where measurements of MSA and nss-$SO_4^=$ are missing.

In the revised manuscript, as a complement to Section 4.1 "Evaluation of ML model performance", we added the following paragraph also adding the Figure to the Supplement materials.

"Knowing that the GPR model could be biased due to the inhomogeneous distribution of in situ observations, we assessed the applicability of the GPR model in regions poorly covered by atmospheric observational data (as the central part of the domain) by running the model in a worst-case scenario deployment. In this exercise, we predicted the daily variations of nss-$SO_4^=$ measurements in the westernmost portion of the study area by training the model only with observations from the eastern part of the domain (i.e., data collected at MHD). In this case, MHD data were used for training/cross-validation, while the four NAAMES campaigns were employed as independent test data. The evaluation on the test data (Fig. S8) reveals that GPR can explain 55% of the daily observed nss-$SO_4^=$ variance (MAE= 0.129 µg m^{-3}), even in this worst-case scenario and on a limited test dataset (n=57). This more than acceptable performance of the model supports the reliability of the IPB-MSA&SO_4 dataset also in the central part of the NA, where measurements of MSA and nss-$SO_4^=$ are missing. In addition, Section 4.5 describes the validation of the GPR model for predicting observed MSA concentrations during the Polarstern campaigns, which were not included in either the model training/cross-validation or in the model test."

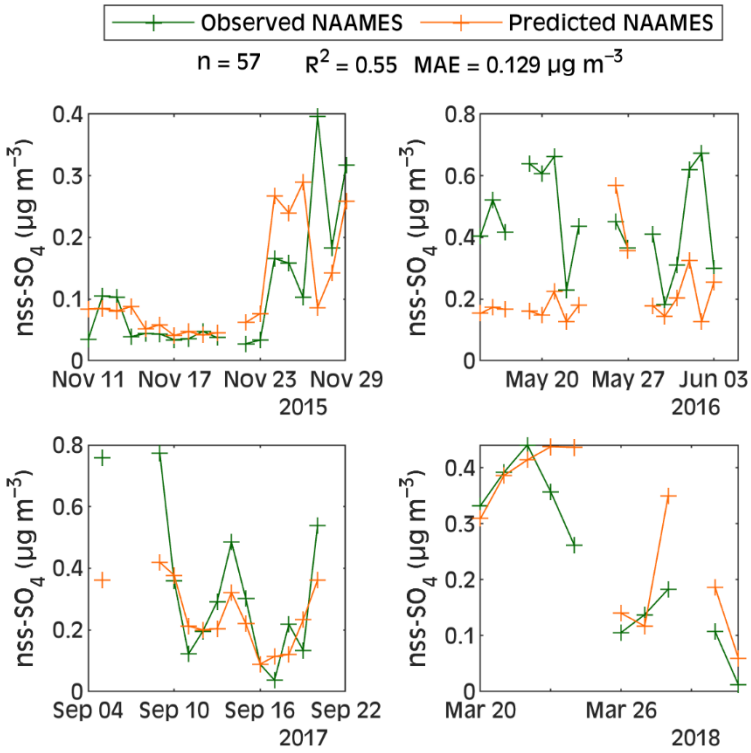


Figure S8 (revised supplementary): Comparison between daily observed and GPR-predicted nss-$SO_4^=$ during the four NAAMES campaigns. The GPR was trained on the MHD data and tested on the NAAMES data. R^2 is computed in a logarithmic space, whereas MAE is computed on a normal scale.

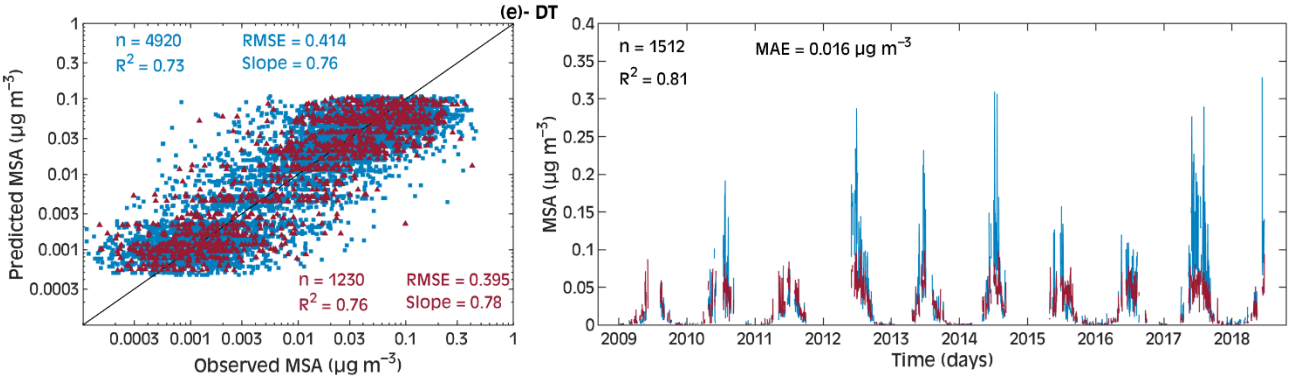
3. You used many variables to train the machine-learning model. However, I felt these predictors were not strong proxy for MSA and sulfate. Why do not use SO2 satellite product for sulfate estimation?

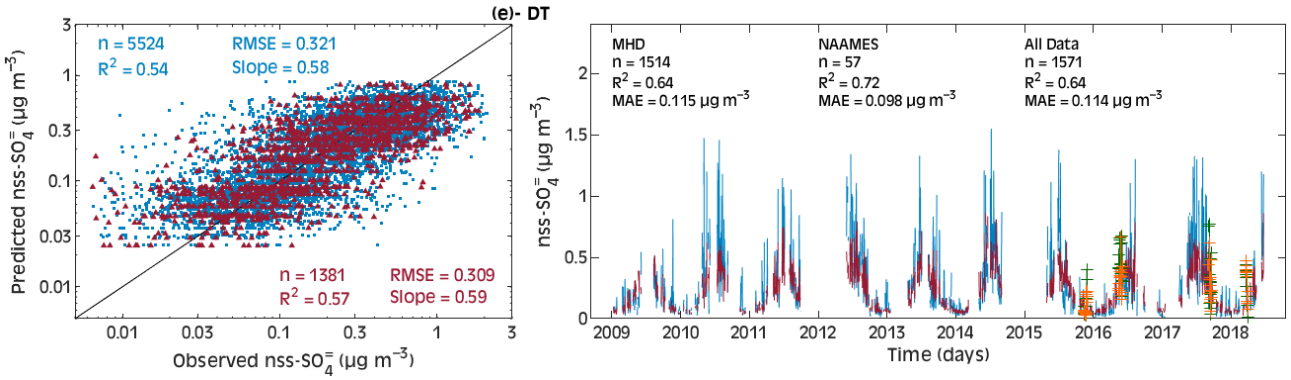
We respectfully disagree with the reviewer's point of view. SO_2 in the atmosphere comes from both natural (e.g., volcanic activity) and anthropogenic sources (industrial processes, fossil fuel combustion by power plants, ships, and other vehicles) with anthropogenic sources contributing importantly at the global level. In this work, we are interested in predicting marine natural biogenic sulfur aerosol concentrations, hence we limit the predictors to variables that can be used as unambiguous tracers of biogenic marine emissions, such as the DMS flux. DMS is the main precursor of MSA and nss-$SO_4^=$ in the marine boundary layer according to at least 40 years of literature, while SO_2 observations may be biased by anthropogenic or volcanic inputs.

The goodness of our choice of predictors is proved by our results: the machine learning-trained models using the selected predictors have very good predictive skills, accounting for as much as 86% and 72% of the daily MSA and nss-SO4$^{=}$ variances (R^2), respectively.

4. Why do you only use the four machine-learning models to predict MSA and sulfate? Please explain the reason. To the best of my knowledge, decision tree model and deep learning might show the better performance compared with ANN and SVM.

We refer to Section 3.2 (in the revised manuscript) where we clarified that the most common types of ML algorithms have been trained under different advanced options and optimizations which can increase the performance and resilience of the algorithms. Following the reviewer's suggestion, we extended the ML models to include the Decision Tree (DT) type. Even, after considering this new ML model, the best performing model is still the GPR; indeed, the DT algorithm provided the lowest performance on our dataset. The panels below have been added to Figure 3 (for MSA) and Figure 4 (for nss-$SO_4^{=}$), respectively.





The following paragraph has been inserted in the revised manuscript and the subheadings numbers and the text have been modified.

"3.2.2 Decision Tree (DT)

The DT model is a non-parametric, non-linear model that generates a structure resembling a tree for classification and regression (Kotsiantis, 2013; Quinlan, 1986). It repeatedly divides the dataset into smaller subsets based on independent features from the input dataset. The split seeks to reduce variability within each group while increasing the variance between subsets. The final tree is made up of decision and leaf nodes. The decision node represents a condition on an attribute, and its branches indicate the conditions' outcomes. For additional information on DT, the reader is directed to https://www.mathworks.com/help/stats/fitrtree.html. The critical parameter in this technique is determining when to terminate the dividing process. In this study, we set up three different minimum leaf sizes (minimum samples to split) to control the number of data that should be in the sub-branch to continue the splitting process, namely 4 (fine tree), 12 (medium tree), and 36 (coarse tree) as seen in Table 1."

We inserted a new table (Table 1) to summarize the ML models used which could be immediate for the readers. Tables S1 and S2 have been merged as one table and present the evaluation measures, accordingly. Lastly, we refer to the use of neural networks in the manuscript which represents the deep learning models. Deep learning is a subtype of machine learning that resembles a neural network with three or more layers.

Model Type	*Preset*	*Hyperparameters if any*
Support Vector Machines	*Linear*	
	Quadratic	
	Cubic	
	Fine Gaussian	*Kernel scale = 0.61*
	Medium Gaussian	*Kernel scale = 2.4*
	Coarse Gaussian	*Kernel scale = 9.8*
Decision Tree	*Fine*	*Minimum leaf size = 4*
	Medium	*Minimum leaf size = 12*
	Coarse	*Minimum leaf size = 36*
Regression Ensemble	*Boosted*	*Minimum leaf size = 8* *Number of learners = 30*
	Bagged	*Minimum leaf size = 8* *Number of learners = 30*
Gaussian Process Regression	*Squared Exponential*	
	Matern 5/2	
	Exponential	
	Rational Quadratic	
Neural Networks	*Narrow*	*Number of fully connected layers = 1* *First layer size = 10*
	Medium	*Number of fully connected layers = 1* *First layer size = 25*
	Wide	*Number of fully connected layers = 1* *1st layer size = 100*
	Bi-layered	*Number of fully connected layers = 2* *1st layer size = 10* *2nd layer size = 10*
	Tri-layered	*Number of fully connected layers = 3* *1st layer size = 10* *2nd layer size = 10* *3rd layer size = 10*

Table 1: List of machine learning models used in the present study.

5. Section 4.6, I think the discussion about the spatiotemporal variations of MSA and sulfate seems to be very superficial and I suggest the authors should add more in-depth analysis.

We appreciate the reviewer's suggestion. In fact, we are focusing on many aspects of the proposed dataset that we believe will provide interesting scientific findings in future publications. Nevertheless, we necessarily need to limit the data exploitation in the present manuscript to avoid exceeding the scopes of ESSD, which is a data journal.

Anyway, in the revised version, we extended the analysis of this section to include more detailed information about the data distribution. The main modifications are:

- The Section title has been renamed to "Spatial distributions of MSA and nss-$SO_4^=$" instead of "Monthly MSA and SO_4 distributions"
- The annual spatial distributions (Figure 8 in the revised manuscript) have been added and the main spatial features has been explained.
- Figure 8 and Figure 9 (old version) have been merged as one figure (Figure 9 in the revised manuscript), to better and immediately compare the monthly variations.
- A new table (Table S2 in the revised supplementary) summarizing the statistics of the annual and monthly climatology (1998-2022) of MSA, nss-$SO_4^=$ and MSA:nss-$SO_4^=$ has been inserted.
- A new figure has been inserted (Figure S10 in the revised supplementary). It presents the spatial distribution of the monthly coefficient of variation (COV) calculated as the percentage of standard deviation divided by the mean, to evaluate the monthly stability of MSA and nss-$SO_4^=$. Higher COV indicates lower stability (many more variants).
- Accordingly, the main text has been modified to include the new analyses.

We refer to the tracked version of the revised manuscript where the modifications have been evidenced.

References

Kotsiantis SB. Decision trees: a recent overview. Artificial Intelligence Review 2013; 39: 261-283.

Mansour K, Decesari S, Ceburnis D, Ovadnevaite J, Rinaldi M. Machine learning for prediction of daily sea surface dimethylsulfide concentration and emission flux over the North Atlantic Ocean (1998–2021) Science of The Total Environment 2023; 871.

Quinlan JR. Induction of decision trees. Machine Learning 1986; 1: 81-106.

Zhou SQ, Chen Y, Wang FH, Bao Y, Ding XP, Xu ZJ. Assessing the Intensity of Marine Biogenic Influence on the Lower Atmosphere: An Insight into the Distribution of Marine Biogenic Aerosols over the Eastern China Seas. Environmental Science & Technology 2023.