IPB-MSA&SO_4: a daily 0.25° resolution dataset of In-situ Produced Biogenic Methanesulfonic Acid and Sulfate over the North Atlantic during 1998–2022 based on machine learning

Karam Mansour[1,2], Stefano Decesari[1], Darius Ceburnis[3], Jurgita Ovadnevaite[3], Lynn M. Russell[4], Marco Paglione[1], Laurent Poulain[5], Shan Huang[5,*], Colin O'Dowd[3] and Matteo Rinaldi[1]

[1] Italian National Research Council, Institute of Atmospheric Sciences and Climate (CNR-ISAC), Bologna 40129, Italy
[2] Oceanography Department, Faculty of Science, Alexandria University, Alexandria 21500, Egypt
[3] School of Natural Sciences, Ryan Institute Centre for Climate and Air Pollution Studies, University of Galway, Galway, Ireland
[4] Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA, USA
[5] Leibniz Institute for Tropospheric Research, Leipzig, Sachsen, 04318, Germany
[*] Now at Institute for Environmental and Climate Research (ECI), Jinan University, Guangzhou, China

Correspondence to: Karam Mansour (k.mansour@isac.cnr.it) & Matteo Rinaldi (m.rinaldi@isac.cnr.it)

Abstract.

Accurate long-term marine-derived biogenic sulfur aerosol concentrations at high spatial and temporal resolutions are critical for a wide range of studies including climatology, trend analysis, model evaluation, accurate investigation of their contribution to aerosol burden, or to elucidate their radiative impacts and to provide boundary conditions for regional models. By applying machine learning algorithms, we constructed the first, publicly available, daily gridded dataset of in-situ produced biogenic methanesulfonic acid (MSA) and non-sea-salt sulfate (nss-$SO_4^=$ ~~SO_4~~) concentrations covering the North Atlantic Ocean. The dataset is of high spatial resolution of 0.25° × 0.25°, spanning 25 years (1998–2022), far exceeding what observations alone could achieve both space- and time-wise. The machine learning models were generated by combining in-situ observations of sulfur aerosol data at Mace Head research station, west coast of Ireland, and from NAAMES cruises in the NW Atlantic, ~~combined~~ with the constructed sea-to-air dimethylsulfide flux (F_{DMS}) and ECMWF-ERA5 reanalysis datasets. To determine the optimal method for regression, we employed ~~four~~ five machine learning model types: support vector machines, decision tree, regression ensemble, Gaussian process, and artificial neural networks. A comparison of the mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R^2) revealed that the Gaussian process regression (GPR) was the most effective algorithm, outperforming the other models in simulating the biogenic MSA and nss-$SO_4^=$ ~~SO_4~~ concentrations. For predicting daily MSA (nss-$SO_4^=$ ~~SO_4~~), GPR displayed the highest R^2 value of 0.86 (0.72) and the lowest MAE of 0.014 (0.10) μg m^{-3}. The GPR partial dependence analysis suggests that the relationships between predictors and MSA and nss-$SO_4^=$ ~~SO_4~~ concentrations are complex rather than linear. Using the GPR algorithm, we produced a high-resolution daily dataset of In-situ Produced Biogenic MSA and nss-$SO_4^=$ ~~SO_4~~ sea-level

concentrations over the North Atlantic, which we named IPB-MSA&SO_4. The obtained IPB-MSA&SO_4 data allowed us to analyze the spatiotemporal patterns of MSA, nss-$SO_4^{=}$~~SO_4~~, and the ratio between them (MSA:nss-$SO_4^{=}$~~SO_4~~). A comparison with the existing CAMS-EAC4 reanalysis suggested~~s~~ that our high-resolution dataset reproduces with high accuracy the spatial and temporal patterns of the biogenic sulfur aerosol concentration and has high consistency with independent measurements in the Atlantic Ocean. The IPB-MSA&SO_4 is publicly available at https://doi.org/10.17632/j8bzd5dvpx.1 (Mansour et al., 2023b).

1 Introduction

Marine-derived biogenic sulfur aerosol particles exert an important influence on the radiative properties of the atmosphere, both directly by scattering solar radiation and indirectly by modifying cloud properties (Langmann et al., 2008; Charlson et al., 1987). Dimethylsulfide (DMS), a volatile organic compound produced by marine microbes~~phytoplankton~~, is the main precursor of biogenic sulfur-containing aerosols in the marine boundary layer (MBL). After being ventilated into the atmosphere, DMS is oxidized to form two of the major secondary marine aerosol species, Methanesulfonic acid (MSA) and non-sea-salt sulfate (nss-$SO_4^{=}$). ~~Throughout the present study, we abbreviate the nss-SO_4^{2-} concentration as SO_4 and MSA concentration as MSA, for simplicity.~~ Sulfur emitted by marine organisms constitutes 20% (Fiddes et al., 2018) to 40% (Simo, 2001) of the total sulfur burden of the atmosphere. The understanding of the role of MSA and nss-$SO_4^{=}$~~SO_4~~ concentrations in Earth's climate is elusive (Mansour et al., 2020a; Hodshire et al., 2019). According to the CLAW hypothesis (Charlson et al., 1987), negative climate feedback is expected to occur if phytoplankton responds to elevated temperature ~~and~~ or solar radiation levels by increasing their DMS production, thereby, exerting a cooling effect by increasing the planetary albedo. Indeed, studies confirmed that DMS emissions contribute significantly to stabilizing the Earth's atmosphere (Sanchez et al., 2018; Thomas et al., 2010; Kim et al., 2018; Mahmood et al., 2019; Mansour et al., 2022; Mansour et al., 2020b), while a few others have claimed that the biological control over cloud condensation nuclei (CCN) goes even beyond the CLAW's climatic feedback role of DMS (Quinn and Bates, 2011; Woodhouse et al., 2010; O'dowd et al., 2004). As a result, biogenic sulfur aerosols play a central role in ocean-atmosphere interactions and regional climate change, and it is critical to parameterize and characterize biogenic MSA and nss-$SO_4^{=}$~~SO_4~~ across different sea areas to constrain the past, current and future climate impacts of both species (Hodshire et al., 2019; Gondwe et al., 2003).

The global aerosol-chemistry-climate general circulation models are used widely to assess the radiative forcing of DMS-derived aerosols. A negative forcing caused by the DMS effect is predicted ranging between −1.7 and −2.3 W m^{-2} (Fiddes et al., 2018; Fung et al., 2022; Thomas et al., 2010; Mahajan et al., 2015). This range is comparable to the positive forcing impact of anthropogenic CO_2 emissions (1.83±0.2 W m^{-2}) (Etminan et al., 2016). Large uncertainties in DMS forcing estimates (up to ±10 W m^{-2}) are partly because models overlook the high-frequency spatial, temporal, and seasonal variability in DMS fluxes (Mansour et al., 2023a; Royer et al., 2015; Mcnabb and Tortell, 2022), and consequent oxidation

products (Riccobono et al., 2014), which are not adequately constrained by the available sparse observations (Bock et al., 2021). This level of uncertainty underlines the need for improved parameterizations of natural sulfur aerosol cycling and fluxes at regional scales (Hulswar et al., 2022; Gali et al., 2018; Mahajan et al., 2015), which is essential for determining their impact on climate. 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.

Focusing on the North Atlantic (NA) Ocean, sulfur-containing aerosols, MSA and nss-$SO_4^=$~~SO_4~~, have been measured at Mace Head sampling station, a coastal area in the eastern NA Ocean, to quantify the contribution of phytoplankton emissions to aerosol mass concentrations in MBL (Rinaldi et al., 2010; Rinaldi et al., 2009; O'dowd et al., 2004), to assess the long-term seasonal patterns in the chemical composition of submicron aerosol in the different origin of marine air masses (Ovadnevaite et al., 2014), and to identify the oceanic regions acting as the main source of biogenic aerosols (Mansour et al., 2020b). During NAAMES field campaigns, research cruises aimed at comprehending the relationships between ecosystems, aerosols, and clouds (Behrenfeld et al., 2019), Saliba et al. (2020) evaluated the origins and contributions of submicron organic and sulfate components to CCN concentrations in the MBL. They concluded that the DMS-derived secondary nss-$SO_4^=$~~SO_4~~ enhanced hygroscopicity, particle size, and CCN concentrations by 5–66%, especially in the spring, highlighting the importance of phytoplankton produced DMS emissions for the CCN budget in the NA (Mansour et al., 2022; Mansour et al., 2020b; Sanchez et al., 2018). However, it is currently challenging to effectively investigate climatology, long-term trends and climate forcing of biogenic sulfur compounds, as well as validate inherent model outputs, since there is a lack of high-time resolution data on these compounds.

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 nss-Sulfate (IPB-MSA&SO_4) 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 NA Ocean, by exploiting the potential of machine learning. The dataset represents the sea-level concentrations of MSA and nss-$SO_4^=$, in each grid point of the domain, resulting from the interplay between precursor emissions and local atmospheric conditions. We created the IPB-MSA&SO_4 dataset using in-situ MSA and nss-$SO_4^=$~~SO_4~~ data measured at Mace Head (MHD) site and from NAAMES cruises, the gridded dataset from the ECMWF-ERA5 together with the reconstructed~~constructed~~ F_{DMS} (Mansour et al., 2023a) as input data. To achieve this aim, we employed machine learning (ML) approaches: support vector machines (SVM), decision tree (DT), regression ensemble (RE), Gaussian process regression (GPR), and artificial neural networks (ANN). ML has been applied in a variety of scientific areas for model approximation, experiment design, and multivariate regression of oceanic and atmospheric complex systems, however, no prior applications to MSA and nss-$SO_4^=$~~SO_4~~ prediction have been published, to our

knowledge. During model training, we evaluated the various possible kernel functions and hyperparameters in each ML type (details in Table S1), employing the 5-fold cross-validation strategy to select the best-performing (optimal) function capable of properly predicting MSA and nss-$SO_4^=$ SO_4. The partial dependence analysis is also used to assess the effect of different predictors on the modeled MSA and nss-$SO_4^=$ SO_4. Furthermore, we investigate the annual and monthly spatial distributions of MSA, nss-$SO_4^=$ SO_4 and the ratio between them (MSA:nss-$SO_4^=$ SO_4) to examine the monthly evolution of MSA and nss-$SO_4^=$ SO_4 in the different regions of the NA domain from 1998 to 2022. The output data (IPB-MSA&SO_4) from this study should be useful for filling the data gap, particularly for the NA, and be applicable to a variety of investigations, such as climatology, trend analysis, model evaluation, radiative impacts, and providing boundary conditions for regional models.

2 Study domain and data sources

2.1 Study area and measuring sites

The study area extends from 20° to 66° N and from 72° W to the prime meridian (Fig. 1) covering the NA Ocean. The key climate-relevant features in the study domain are the Gulf Strem, its northern extension towards Europe known as the North Atlantic Current (NAC), Atlantic meridional overturning circulation (AMOC) (Buckley and Marshall, 2016) and the cyclonic subpolar gyre (SPG) (Rhein et al., 2011). The Gulf Stream is a warm Atlantic Ocean flow that begins in the Gulf of Mexico and moves through the Straits of Florida before continuing up the eastern coast of the United States (Buckley and Marshall, 2016). These warm northward-flowing waters meet the cold southward-flowing waters of the Labrador Current and the western boundary current of the cyclonic subpolar gyre, ultimately turning east and heading toward Northwest Europe as the NAC. The NAC then splits into multiple branches that enter the subpolar gyre, one of which passes via the Iceland Basin and the other through the Rockall trough (Fratantoni, 2001). AMOC is a major current system of the NA transporting the warm and salty surface waters toward the North and the cold deep waters toward the South. The NA SPG extends from 45° N to around 65° N and comprises the sills between Greenland, Iceland, the Faroe Islands, and Scotland. Such circulation phenomena areThe SPG is a crucial region for the modulation of the temperate climate of north-western Europe (Marzocchi et al., 2015), and its the dynamics of SPG determine the rate of deep and intermediate water formation (sinking dense and cold surface waters through air-sea heat exchanges in the wintertime) particularly in the Labrador Sea (Katsman et al., 2004). Both phenomena Accordingly, they contribute to the regional changes in of primary production biological activity and the subsequent biogenic emissions in the study domain.

The MHD global atmospheric watch (GAW) research station (53.33° N, 09.90° W) is located on Ireland's west coast (Fig. 1), at about 80 meters from the coastline and 21 m above mean sea level. MHD is the only GAW station in the eastern Atlantic region and is the globally acknowledged clean background western European station, providing key baseline input for intercomparing with levels elsewhere in Europe (Grigas et al., 2017; O'dowd et al., 2014).

Four shipboard field campaigns were carried out as part of the NAAMES research project (Behrenfeld et al., 2019). The tracks of cruises representing marine conditions during aerosol sampling (Saliba et al., 2020) are shown in Fig. 1. The measurements cover the periods of November 2015, May–June 2016, September 2017, and March 2018. Behrenfeld et al. (2019) provide a thorough explanation of the NAAMES project's goals, objectives, and atmospheric and oceanic conditions.

2.2 Observational data

The long-term submicron sulfur aerosol species atmospheric concentrations (Methanesulfonic acid [MSA] and Sulfate [$SO_4^{=}$ ~~SO_4~~]) from January 2009 to June 2018 measured at MHD were used. The measurements were performed by using the Aerodyne High Resolution- Time of Flight- Aerosol Mass Spectrometer (HR-ToF-AMS). The HR-ToF-AMS (Decarlo et al., 2006) output has a time resolution of ~5-10 minutes and it was operated according to the recommendations by Jimenez et al. (2003), Allan et al. (2003) and Canagaratna et al. (2007). The MSA was derived from the concentration of mass fragment $CH_3SO_2^+$ (Ovadnevaite et al., 2014). Further information on the MSA measurement can be found in Mansour et al. (2020a). The black carbon (BC) concentrations were measured in-situ at MHD by a multi-angle absorption photometer (O'dowd et al., 2014) to identify the anthropogenically impacted air masses, as detailed in Section 3.1.1.

High-resolution in-situ shipborne measurements of non-refractory submicron $SO_4^{=}$ ~~SO_4~~ concentrations were measured every 5 min using HR-ToF-AMS during four open-ocean research cruises (NAAMES) in the NW Atlantic [4 campaigns represent winter (November 2015), late spring (May–June 2016), autumn (September 2017), and early spring (March 2018)] (Saliba et al., 2020). We employ the $SO_4^{=}$ ~~SO_4~~ concentrations, whereas there are no high-resolution MSA datasets available from NAAMES campaigns, during periods that were largely marine aerosol sources which were defined as periods when particle number concentrations <1500 cm^{-3}, BC <50 ng m^{-3}, 2-days back trajectories originated from the North or tropical Atlantic, and radon concentrations <500 mBq m^{-3} according to Saliba et al. (2020). The measured $SO_4^{=}$ ~~SO_4~~ from AMS excludes refractory particles that likely contain the majority of sea-salt sulfate which is therefore approximately equivalent to nss- ~~sulfate~~ $SO_4^{=}$ (Frossard et al., 2014).

2.3 Air mass back-trajectories

The Air Resources Laboratory (ARL) of the National Oceanic and Atmospheric Administration (NOAA) developed the Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT4) model (Rolph et al., 2017; Stein et al., 2015), which is used to calculate the air mass back-trajectories (BTs). The archived Global Data Assimilation System (GDAS1) (1° × 1°) of the National Centers for Environmental Prediction (NCEP) was used as a driver of the trajectory calculation (ftp://arlftp.arlhq.noaa.gov/pub/archives/gdas1). We run the model at the MHD sampling station as a fixed source location and throughout the NAAMES cruises as a moving source location. The starting height is set to be 100 m above ground level and the backward time is 3 days with an interval of 1 h along each entire trajectory track. The schematic diagram of BTs calculation is shown in Fig. S1. The arrival frequency of BTs at MHD is 3h (eight tracks a day) covering the period from 01-

Jan-2009 to 30-Jun-2018 and of NAAMES is hourly (twenty-four tracks a day) covering the time of the four campaigns identified as marine periods (Saliba et al., 2020).

2.4 Dimethylsulfide flux data

The seawater DMS is the primary contributor to biogenic sulfur aerosol in the atmosphere. For this reason, we use the sea-to-air DMS flux (F_{DMS}) as a predictor of MSA and nss-$SO_4^=$ ~~SO_4~~ concentrations. Mansour et al. (2023a) used an ML predictive algorithm based on Gaussian process regression (GPR) to simulate the distribution of daily seawater DMS concentrations and related F_{DMS} in the NA areas from 35° to 66° N and from 0° to 55° W at 0.25° × 0.25° spatial resolution. We extended the GPR model within the NA to encompass the NAAMES measurements, which are essential because they cover the western most section of the study area. Fig. S2 displays the main differences between the two domains. Simply, the GPR was trained once more, utilizing the same approach of Mansour et al. (2023a), with a higher number of data points and yielded an enhanced R^2 value up to 0.77 on the independent test dataset. The daily sea-to-air F_{DMS} was calculated using the gas transfer velocity (Goddijn-Murphy et al., 2012) and the DMS derived from GPR predictions. For more details about the data product, we refer the reader to Mansour et al. (2023a).

2.5 Meteorological data

The ECMWF-ERA5 reanalysis data (Hersbach et al., 2020) were downloaded to extract the meteorological parameters used as predictors of MSA and nss-$SO_4^=$ ~~SO_4~~ in the ML models. ERA5 provides estimates for the hourly state of the atmosphere, worldwide, with spatial resolution 0.25° × 0.25° at the surface and different pressure levels. From the global domain, we extracted multiple atmospheric components including air temperature at 2m above sea level (AT) and surface net short-wave radiation flux (SRF) as representative of thermal heating, and the relative humidity (RH) as representative of water vapor abundance in the atmosphere. To represent the dispersion of aerosol particles in the troposphere and the wet removal through the below-cloud scavenging process, the boundary layer height (BLH) and the precipitation rate (PR) were utilized, respectively.

3 Methods

3.1 Data preparation

In this Section, we describe the preparation of predictors and responses that were used to train, cross-validate, and generate the ML models.

3.1.1 Air mass selection

In previous studies (Mansour et al., 2020b; O'dowd et al., 2015; Ovadnevaite et al., 2014), BC concentration was often considered as a useful tool to select clean marine air masses excluding inputs from continental emissions or ship trails. In this study, we still relied on BC measurements as a precious tool to identify and exclude anthropogenically impacted air masses, but we also developed a more complete approach aimed at identifying air masses characterized by a high degree of contact with the ocean surface. This was necessary in order to select, from the in-situ observations, data points representing almost entirely oceanic sources to provide the best dataset for training the ML models.

The retention ratio of the air mass over the ocean (R_O) was calculated to determine whether an air mass (identified by BT track) arriving at the MHD sampling station or at the ship location, in the case of shipborne measurements, was primarily from the NA region or not. We used 3-day BTs arriving 100 m above the MHD sampling station and NAAMES tracks. The BTs tracks at the MHD arrival point were calculated 8 times per day, whereas it was 24 times per day at NAAMES measuring points, considering only the measurements classified as marine periods (Saliba et al., 2020). The R_O has been calculated for each track as:

$$R_O = \frac{\sum_{i=1}^{N_{Ocean}} e^{\frac{-t_i}{72}}}{\sum_{i=1}^{N_{Total}} e^{\frac{-t_i}{72}}} \quad (1)$$

where N_{Total} is the total number of trajectory endpoints which is equal to 73 (arrival point + 72 backward hours). N_{Ocean} is the total number of trajectory endpoints passing over the ocean, while t_i is the backward tracking time with the unit of an hour spanning the values from 0 to 72. Because air mass diffusion and particles deposition potentially occur during the air mass transport, a weighting factor $e^{-t_i/72}$ related to tracking time has been introduced. The weighting factor takes the values from 1 (at the arrival point) up to 0.37 (farthest point), hence, the oceanic areas far from the arrival point, corresponding to longer backward tracking time, have a weaker influence than areas closer to the sampling point. As a result, a higher R_O value implies that oceanic emissions have a greater influence on the air mass and that the source region is more likely to be the ocean. Other studies have used similar methods to characterize air mass source regions. For example, Zhou et al. (2021) studied the contribution of non-marine MSA sources in the coastal East China Sea and the Gulf of Aqaba by characterizing the land air masses. Rinaldi et al. (2021) used a combination of low-travelling air mass BTs and satellite ground-type maps to investigate the effect of ground conditions (sea ice, snow, seawater, and land) on air samples at Ny-Ålesund station in the Arctic Ocean.

Because oceanic air masses crossing the NA can pass above the BLH, its connection to local sea surface processes such as marine biogenic emission and subsequent atmospheric reactions may be significantly weaker. To address this issue, Eq. 2 was used to calculate the retention ratio of an ocean air mass within the marine boundary layer (R_B).

$$R_B = \frac{\sum_{i=1}^{N_{Below}} e^{\frac{-t_i}{72}}}{\sum_{i=1}^{N_{Ocean}} e^{\frac{-t_i}{72}}} \quad (2)$$

where N_{Ocean} is the total number of trajectory endpoints located over the ocean (i.e., marine endpoints) and N_{Below} is the number of marine endpoints which have an altitude below BLH. The higher the R_B value, the more airflow over the ocean is confined to the MBL. The BLH datasets at each endpoint were extracted from the hourly ERA5 dataset.

The total number of BTs tracks arriving at MHD during the period from Jan-2009 to Jun-2018 is 27,744 (3468 days × 8 tracks per day). We counted the number of endpoints of all BTs in each 1° × 1° grid cell and normalized them to the maximum value to find the percentage of endpoints for all grid cells (Fig. S3). The larger density of BTs endpoints is concentrated over the NA oceanic region, indicating that the main source regions for air masses transported to MHD sampling stations are most likely oceanic. At MHD, we investigated how MSA (a marine biogenic tracer) responds to change in BC (a tracer of anthropogenic input) as seen in Fig. S4, by considering hourly data simultaneous to the arrival time of BTs (*i.e.*, 8 times a day). We found that MSA tends to fluctuate minimally when BC is less than 15 ng m^{-3} (slope = 0.05), whereas MSA tends to rise slightly when BC exceeds 15 ng m^{-3} (slope = 0.28). Such cases with hourly BC concentrations <15 ng m^{-3} were classified as representative of marine conditions, that are likely not influenced by anthropogenic sources. To constrain the impact of marine biogenic emissions and meteorological parameters on MSA and nss-$SO_4^=$~~SO4~~, air masses were included in this analysis only if they were characterized by $R_O + R_B \geq 1.75$, meaning that the air mass had a high degree of contact with the ocean surface within the last 3 days (Fig. S4). Indeed, considering the above condition, an air mass must have at least R_O equal to 0.75 and in such case the track must be traveling 100% of the time below the BLH. By introducing the criterion of $R_O + R_B \geq 1.75$, approximately 72% of the BTs tracks were considered. This reflects the significance of the MHD research station for studying NA biogenic emissions, and the frequency with which it is impacted by MBL air masses (Grigas et al., 2017; O'dowd et al., 2014). After considering the BC threshold (<15 ng m^{-3}) and conservatively removing all the observations done when the BC data were unavailable (instrument downtime), 9211 (33% of the total) tracks were classified as representative of marine conditions (selected marine BTs frequency is presented in Fig. S5).

Regarding the NAAMES measurements, the total number of calculated BTs tracks was 832 (Fig. S6) during background marine conditions, identified by Saliba et al. (2020). In this study, we kept 660 tracks (Fig. S7) of the above 832 as representative samples of marine conditions during NAAME~~A~~S cruises by limiting the analysis to hourly samples with $R_O + R_B \geq 1.75$.

3.1.2 Predictors extraction along back trajectories

In order to train the ML models, it was necessary to associate each observed MSA and nss-$SO_4^{=}$ ~~SO_4~~ data point with the corresponding potential predictors. The potential predictors (F_{DMS}, AT, SRF, RH, BLH and PR) were extracted at each endpoint of the BTs associated with each of the selected clean marine observational data points (see Section 3.1.1), inside the oceanic region within 20−66 °N and 0−72 °W (Fig. S1). The extracted predictor values were then averaged along each marine BT track, providing the most representative picture of the conditions (air mass history) that led to the formation of the observed sulfur aerosol concentrations. The few endpoints over land or crossing above the BLH were eliminated.

The Pearson's correlation coefficients between the potential predictors and observational MSA and nss-$SO_4^{=}$ ~~SO_4~~ data were compared, considering different BT lengths of 1, 2 and 3 days, to assess which BT length was more representative of the time scale of sulfur aerosol formation processes. As seen from Table ~~1~~2, both MSA and nss-$SO_4^{=}$ ~~SO_4~~ correlate better with F_{DMS} considering a 3-day BT length. Similarly, the majority of the other predictors, except for AT, tended to maximize their correlations considering 2 or 3 days of BT length. Ultimately, we considered for each predictor the BT length that maximized the correlation coefficient for the analyses in the present study.

3.1.3 Responses at measuring sites

Hourly nss-$SO_4^{=}$ ~~SO_4~~ at MHD and from NAAMES campaigns as well as MSA at MHD, measured concurrently with the selected marine BTs (Section 3.3.1), were used to build ML models. A total of 6162 (6920) data points for MSA (nss-$SO_4^{=}$ ~~SO_4~~) were obtained. Further, we also applied 0.1 and 99.9 percentiles lower and upper thresholds filter to remove the extremely low and high values that could bias the ML models training and cross/validation. This helped to identify and remove outliers in each dataset, thereby reducing the number of data points to 6150 (6905) for MSA (nss-$SO_4^{=}$ ~~SO_4~~) (~0.2 % of data points were rejected). Details of the MSA and nss-$SO_4^{=}$ ~~SO_4~~ percentile thresholds, along with the amount of data before and after applying the filters are given in Table ~~2~~3. The hourly data after cleanup is used for training/ cross-validation and testing of ML models.

3.2 Machine learning models

The methodological flowchart of the present study is shown in Fig. 2. The core of the framework is using the supervised ML regression techniques to build predictive models for estimating the atmospheric concentrations of biogenic MSA and nss-$SO_4^{=}$ ~~SO_4~~ (responses) from independent variables (predictors). Predictors include the sea-to-air F_{DMS} and meteorological parameters that control the aerosol concentration in the MBL. ~~Given that ML models may be generated even if there is no physical relationship between predictors and responses, w~~We used multilinear regression to assess the contribution of each predictor to MSA and nss-$SO_4^{=}$ ~~SO_4~~ variations. Initially, we ran the multilinear regression model using the total of the potential six predictors: F_{DMS}, AT, SRF, RH, BLH and PR. Secondly, we applied the multilinear regression models by

eliminating one predictor each time. Each independent variable's contribution to R^2 is the reduction in total R^2 when that variable is eliminated. The results (Table ~~3~~4) showed that the six predictors used can explain up to 74% (53%) of MSA (nss-$SO_4^=$~~SO_4~~) variance. Such predictors tend to contribute differently to MSA and nss-$SO_4^=$~~SO_4~~. SRF, F_{DMS} and BLH are the most effective parameters for MSA (explaining up to 64 % of the variability), while SRF, AT and F_{DMS} are the most influential on nss-$SO_4^=$~~SO_4~~ (explaining up to 44 % of the variability). RH has a minor contribution to the MSA and nss-$SO_4^=$~~SO_4~~ variance. To know if a predictor contributes significantly to the explained variance, we performed the analysis of variance (ANOVA) on the implemented multilinear regression model. The ANOVA revealed that all the tested predictors have statistically significant ($p < 0.05$) contributions to MSA and nss-$SO_4^=$~~SO_4~~. For these reasons, we applied the ML models using all of the potential six predictors.

The datasets, containing the corresponding predictors and each one of the responses (MSA and nss-$SO_4^=$~~SO_4~~) separately, were split randomly into two subsets, defined as the training/cross-validation set and the test/evaluation set for each response. The training/cross-validation sets include 80% of the total points (n = 4920 for MSA and n = 5524 for nss-$SO_4^=$~~SO_4~~), while the test/evaluation sets comprise the remaining 20% (n = 1230 for MSA and n = 1381 for nss-$SO_4^=$~~SO_4~~). To improve ML algorithms' accuracy and protect against overfitting, a k-fold cross-validation strategy, with $k = 5$ was used, as this has been shown to provide maximal model prediction robustness and minimal bias (Rodriguez et al., 2010; Fushiki, 2011). The k-fold cross-validation is a procedure used to estimate the skill of the model on new data and generally results in a less biased estimate of the model skill. The number k-fold refers to how many groups a given data sample is to be split into. In this study where $k = 5$, the training/cross-validation dataset randomly was further divided into 5 folds of roughly equal size. At each trial, one group is designated as a holdout or validation dataset, while the remaining four groups are designated as training data (Fig. 2). The model is then fit on the training set (4 folds) and evaluated on the validation set (last fold), and the average evaluation measures (accuracy) on the validation subsets of the five iterations are reported. To better examine the model's repeatability on a new independent dataset, the generated models were evaluated on the test data that was not included in the model construction.

~~Four~~ Five types of ML models were trained/cross-validated and evaluated to identify the best-performing model in estimating sulfur aerosol concentrations (MSA and nss-$SO_4^=$~~SO_4~~). The ML algorithms are SVM, DT, RE, GPR, and ANN. These are the most common types of algorithms, but still, there are subtypes where advanced options and optimizations in the model can increase the performance and resilience of the algorithms. In general, each supervised ML model performs differently and has various strengths and shortcomings. Finding the proper ML algorithm is largely based on trial and error; even experienced data scientists cannot anticipate if an algorithm will work without testing it. Thus, understanding the fundamentals of various ML algorithms and their applicability in diverse applications is critical (Sarker et al., 2019). As a result, initially, we assessed ~~17~~20 algorithms belonging to the aforementioned ~~four~~ five types and chose the most fitted from each type (Table~~s S1 and S2~~ 1), as detailed in the following Sections.

3.2.1 Support vector machines (SVM)

SVM is a powerful mathematical model based on the statistical learning theory (Vapnik, 2013) that can be used either for classification or regression analysis. In recent decades, SVM demonstrated high prediction accuracy in a wide range of regression problems in fields such as oceanography, meteorology, and atmospheric sciences (Lins et al., 2013; Sachindra et al., 2018; Shabani et al., 2020; Shrestha and Shukla, 2015; Fan et al., 2018). The SVM model estimates the regression using a series of kernel functions that are capable of implicitly converting the original, lower-dimensional input data to a higher-dimensional feature space. To achieve the best prediction accuracy for MSA and nss-$SO_4^=$ ~~SO_4~~, we assessed the SVM different kernel functions such as linear, polynomial (quadratic and cubic) and Gaussian (Table ~~S1 and S2~~ 1). The Gaussian kernel was adopted by trying various kernel scales, setting them to 0.61 (fine), 2.4 (medium), and 9.8 (coarse). For more information on SVM, the reader is referred to https://www.mathworks.com/help/stats/fitrsvm.html.

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.

3.2.~~2~~ 3 Regression ensemble (RE)

The ensemble is a technique that employs a collection of DT models (referred to as weak learners or base models), each of which is produced by applying a learning process to a specific problem and then combining them to provide the final prediction (Mendes-Moreira et al., 2012). The performance and accuracy of ensembles are determined by the aggregation of weak learners (Hengl et al., 2018). The well-known types of aggregation are the bagging and boosting methods (Breiman, 2001). In the bagging method (also known as bootstrap aggregating), the base models are generated using random sub-samples drawn from the original dataset with the bootstrap sampling method, where some original examples appear several times while others do not appear at all. On the other hand, the main idea of the boosting method is that it is possible to convert a base model that performs slightly better into one that arbitrarily achieves high accuracy. This conversion is performed by combining the estimations of several predictors. For more information on RE, the reader is referred to https://www.mathworks.com/help/stats/fitrensemble.html.

3.2.~~3~~ 4 Gaussian process regression (GPR)

GPR is a non-parametric technique for solving nonlinear regression problems (Williams and Rasmussen, 1996) which is based on Bayesian theory and statistical learning theory. The accuracy of GPR is dependent on the adopted kernel (covariance) functions (Verrelst et al., 2016). We assessed the different base kernel functions, namely exponential, Matern 5/2, squared exponential, and rational quadratic (Asante-Okyere et al., 2018; Mansour et al., 2023a) to determine the optimal covariance function that could produce reliable predictions of MSA and nss-$SO_4^=$~~SO4~~. For more information on GPR, the reader is referred to Mansour et al. (2023a) and https://www.mathworks.com/help/stats/fitrgp.html.

3.2.~~4~~ 5 Artificial neural networks (ANN)

ANN is an information processing system, which can be used to understand the complex nonlinear relationship between the response and predictors (Kalogirou, 2001). It consists of interconnected groups of artificial neurons that work in the same way as biological neurons. The ANN structure comprises three distinctive groups called input (corresponds to the predictors), several hidden layers (fully connected), and output (corresponds to the predicted response values). The input introduces data to the ANN model, the hidden layer processes the data, and the results are produced in the output. Further details on ANN can be found at https://www.mathworks.com/help/stats/fitrnet.html. We trained various types of ANN as single-layer (number of fully connected layers = 1), bi-layered (number of fully connected layers = 2), and tri-layered (number of fully connected layers = 3) neural networks as detailed in Table~~s S1 and S2~~ 1.

3.3 Evaluation measures

In this study, we use different validation metrics to evaluate the ML models' performance. Each of the metrics is calculated using "residuals". Residuals are the differences between the observed data points O_i and the predicted values P_i, where $i = 1,2,\dots n$. n refers to the number of observations. Better models in predicting the response have residuals close to zero. The average magnitude of the residuals is called mean absolute error (MAE).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|O_i - P_i| \quad (3)$$

Regression models tend to use the square of the residuals instead of the absolute. The square root of the average of the squared residuals is called root mean square errors (RMSE). A low RMSE is a confidence that your model has relatively few large errors.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i}^{n}(P_i - O_i)^2} \quad (4)$$

The metrics listed in Eqn. 3 and Eqn. 4 can only tell you how a model compares to observations and/or other models. Neither can say whether a model is a good fit for the data objectively. Comparing a model to a simple baseline model is a different

approach. This is the motivation behind the use of the coefficient of determination (R^2) metric (Eqn. 5). R^2 is the relative difference in the total error obtained by fitting a model, so a value between 0 and 1. If a model fits the data well, the model error is small and R^2 will be close to 1 and vice versa.

$$R^2 = 1 - \frac{\sum_i^n (O_i - P_i)^2}{\sum_i^n (O_i - \overline{O_i})^2} \quad (5)$$

Where $\overline{O_i}$ is the average of observations. The predicted-observed linear slope is the last metric used to evaluate the performance of ML models. It determines the rate of change of the predicted variable concerning the observed variable and should be close to unity for skilled model predictions.

4. Results and Discussion

4.1 Evaluation of ML model performance

As a first step, we assessed different possible hyperparameters optimization in each type of the ~~four~~ five used ML models (SVM, DT, RE, GPR, and ANN) to determine which one has the best fit and lesser errors in sulfur aerosol (MSA and nss-$SO_4^=$ ~~SO_4~~) predictability. We chose the best model with the least errors in each type for further evaluation and analysis based on the evaluation measures (RMSE, MAE, and R^2). The evaluation measures are summarized in Table S1 ~~for MSA and Table S2 for SO_4~~. The medium Gaussian SVM which utilizes a Gaussian kernel scale equal to the square root of the number of predictors (= 2.4), displayed better performance. The coarse DT, which sets the minimum sample size to split equal to 36, the ensemble bagged trees (EBT) of a bootstrap aggregated ensemble and the GPR, which employs the rational quadratic kernel, represent the minimum errors. Finally, a medium ANN of layer size 25 with one fully connected layer is selected. The ~~four~~ five best-performing (optimal) models have been exported and saved so that they can be used to make new predictions on a new dataset.

Fig. 3a-~~d~~ e and Fig.4a-~~d~~ e present the detailed comparison between observed and predicted MSA and nss-$SO_4^=$ ~~SO_4~~, respectively, of the ~~four~~ five developed ML optimal models. When compared to the multilinear regression (Table ~~3~~4), it is clear that ML models, in general, can reconstruct the observations with a markedly higher R^2 value, which means that the selected ML approaches capture much more of the observed MSA and nss-$SO_4^=$ ~~SO_4~~ variability. While the ~~four~~ five applied optimal algorithms have quasi-similar measures, the best model is GPR for predicting MSA and nss-$SO_4^=$ ~~SO_4~~. For hourly MSA (nss-$SO_4^=$ ~~SO_4~~), the GPR achieves the highest R^2 value of 0.79 (0.64) and the least RMSE of 0.362 (0.282) for the cross-validated data (average measures of each validation fold). When extending to the test data, R^2 and RMSE reach 0.81 (0.67) and 0.347 (0.272), respectively. The EBT comes second in terms of performance in predicting MSA (nss-$SO_4^=$ ~~SO_4~~) with R^2 = 0.80 (0.64) of the independent test data. The SVM and ANN achieve a reasonable accuracy with R^2 = 0.79 (0.61) and 0.78

(0.60), respectively for MSA (nss-$SO_4^=$ ~~SO_4~~) based on the test data. Lastly, based on the hourly test data, the DT shows the lowest, but still respectable, accuracy with $R^2 = 0.76$ for MSA and = 0.57 for nss-$SO_4^=$.

Importantly, the implemented ML models can reconstruct MSA and nss-$SO_4^=$ ~~SO_4~~ daily time series characteristics with remarkable consistency between observed and predicted data. It is worth noting that the daily averages of MSA and nss-$SO_4^=$ ~~SO_4~~ have been calculated from the validation folds and the test set. The MAE of GPR is close to 0.014 (0.100) μg m^{-3} for MSA (nss-$SO_4^=$ ~~SO_4~~). The MAE of EBT, SVM, ~~and~~ ANN and DT are higher than those of both GPR. According to the R^2, the ranking order is the same as for MAE, i.e., GPR outperforms EBT, SVM, ANN and DT ~~and ANN~~ in both MSA and nss-$SO_4^=$ ~~SO_4~~, notwithstanding the differences in the R^2 of the ~~four~~ five models are small. An in-depth look at the MAE and R^2 from MHD and NAAMES (Fig. 4; right panels) demonstrates that the ML models perform well in predicting nss-$SO_4^=$ ~~SO_4~~ across different datasets. All ~~four~~ five models show relatively high values of R^2 on the NAAMES dataset. EBT, SVM and ANN have R^2 values that are similar and equal to 0.81, whilst GPR has ~~a~~ the ~~higher~~ highest value of R^2 reaching 0.87 and DT has the smallest at 0.72. In essence, the performance metrics indicate that GPR always has the highest accuracy and lowest errors, reflecting the robustness of GPR. Therefore, GPR was selected as the optimal regressor for further analysis throughout this study.

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.

4.2 Partial dependence analysis

The bulk of ML models is called a "black box" since the internal computations inside multiple operational layers in a model are concealed and most systems have only observable inputs and outputs out of the box. The partial dependence analysis (Friedman, 2001) is used to assess how predictors influence an output by ML model and show whether the relationship between the response and any of the features is linear, monotonic or more complex. The method entails altering one feature and constraining the remaining features to unaltered average values to illustrate the marginal effect of the changed feature on

the expected outcome. The partial dependence plots of MSA and nss-$SO_4^=$ as a function of the predictors in the highest-performing GPR model are shown in Fig. 5, indicating that the interactions between predictors and response are complex in general. MSA and nss-$SO_4^=$ levels tend to rise as F_{DMS} levels rise from 3 to 10 μmol m^{-2} d^{-1}. MSA continues to rise with stronger F_{DMS} emission rates (>10 μmol m^{-2} d^{-1}), nevertheless, nss-$SO_4^=$ concentration appears independent of F_{DMS} after this threshold. AT exhibits a positive relationship with MSA and nss-$SO_4^=$ concentration in the range of (5–10 °C) and above a downward trend. RH, which has the least impact on MSA and nss-$SO_4^=$ (Table ~~3~~4), has an unclear pattern on the MSA and nss-$SO_4^=$ marginal changes. MSA and nss-$SO_4^=$ present a negative dependence on PR as rain is expected to scavenge aerosol particles; nevertheless, at higher levels of PR, nss-$SO_4^=$ concentrations tend to increase. This may be partly linked to enhanced cloudiness, associated to high PR, where the aqueous phase formation of nss-$SO_4^=$ in the MBL may be favored (Zhu et al., 2006; Von Glasow and Crutzen, 2004). This is also in agreement with the enhancement of nss-$SO_4^=$ concentration at high RH. Finally, BLH and SRF are the most straightforward influencing parameters on MSA and nss-$SO_4^=$ levels, with deep BLH resulting in a dilution of their concentrations and high SRF leading to high MSA and nss-$SO_4^=$ levels, as expected for DMS photo-oxidation products.

4.3 The IPB-MSA&SO_4 dataset

The GPR model was used to generate the long-term gridded fields of high-resolution (0.25° × 0.25°) MSA and nss-$SO_4^=$ concentrations. At each pixel, a daily time series of MSA and nss-$SO_4^=$ have been generated spanning from 1998 to 2022 (9131 days). The total number of pixels in the entire NA domain is 43840, for a total of 400'303'040 data points. The daily time series of MSA and nss-$SO_4^=$ averaged over the entire NA domain are presented in Fig. ~~S8~~S9. The dataset represents the sea-level concentrations of MSA and nss-$SO_4^=$ associated with in-situ production in the MBL derived based on the six selected predictors, which in turn represent the sea-to-air flux of DMS (the precursor) and the meteorological conditions that can mostly affect, in one direction or in the other, the formation of the two products. For this reason, we consider the data to be representative of the concentration of sulfur aerosol species resulting, in each pixel, from the local biogenic emissions in combination with local atmospheric conditions. As such, we called the achieved data product the In-situ Produced Biogenic MSA and nss-$SO_4^=$ (IPB-MSA&SO_4) dataset across the NA. It is important to note that atmospheric motion is not considered in our product and that the maps resulting from the data represent a static picture of potential sea-level concentrations of MSA and nss-$SO_4^=$, in a certain pixel and at a certain time as a result only of the interplay between local DMS emissions, photochemistry and dilution/removal processes, and that provide accurate predictions of the actual sea level concentrations of MSA and nss-$SO_4^=$ once averaged over 2-to-3-days transport tracts. Accordingly, the IPB-MSA&SO_4 data presented hereafter are different from the output of a chemical transport model. Nevertheless, we believe that this unprecedented dataset may be useful for many research purposes, for instance, investigating long-term trends, or addressing the interannual or spatial variability in the production of biogenic sulfur aerosol species. Examples of the scientific

information that can be extracted from the data and on how they can be compared to model output or in-situ observations are provided in the next Sections.

4.4 Comparison with CAMS Reanalysis

To further examine the effectiveness of our GPR model, we compared the observed MSA concentrations at MHD with the most recently released CAMS-EAC4 (Inness et al., 2019) reanalysis datasets. The EAC4 (ECMWF Atmospheric Composition Reanalysis 4) is the fourth generation of the ECMWF global reanalysis dataset of atmospheric composition from the Copernicus Atmosphere Monitoring Service (CAMS). CAMS-EAC4 is a collection of atmospheric composition fields from 2003 to the present, including aerosols and chemical species for which MSA data is available. The spatial resolution of the CAMS datasets is about 0.75° × 0.75° and a 3h temporal resolution. Our datasets have a resolution of 0.25° × 0.25° and start from 1998. To compare the two products, we extracted MSA data from CAMS locally, at the grid cell in front of the MHD station, corresponding to maritime BT timings, and averaged them to daily resolution. Conservatively, the MSA concentration data simulated by GPR were taken from the validation and test sets, which were not included in the model training. Such MSA concentrations at MHD were projected by incorporating predictors along the BTs into consideration to account for the air motion (see Section 3.1.2 for details).

Scatter plots and joint probability histograms of residual errors (Fig. 6) were constructed to compare the accuracy between GPR, CAMS and observations (referred to as OBS). It can be seen from the scatter plots (Fig. 6a and Fig. 6b) that the GPR-simulated MSA best matches the observations, with a ~~1.03~~0.84 fitted slope, 0.93 correlation coefficient and most of the data points comprised within the 95% confidence bounds. The joint probability histograms between observed MSA and the residuals (OBS – GPR) and (OBS – CAMS) are used to verify the variance of residual errors around zero. The GPR histograms (Fig. 6c and Fig. 6e) show that the residual errors are mostly centered around zero (dashed black line in the right) up to the value of 0.1 μg m^{-3} where the majority of data points lie, while CAMS are skewed toward negative residuals followed by positive residuals mainly at high MSA values (Fig. 6d and Fig. 6f). Quantitatively~~Quantitively~~, the GPR has relative MAE equal to 4.3% in comparison to 6.3% for CAMS. In summary, GPR better captures the low concentrations of MSA, which CAMS tends to overestimate, while both CAMS and GPR show limitations in retrieving the extreme points of MSA concentrations. A quantitative statistical analysis (Fig. 6g) showed that no statistically significant ($p<0.05$) difference exists between the seasonal median MSA from OBS and GPR, while CAMS presents a significant ($p<0.05$) difference in all seasons except summer. Nevertheless, the two datasets (GPR and CAMS) properly retrieve the observed MSA seasonal cycle.

4.5 Comparison with the Polarstern cruise results

In this Section, we present a case study exemplifying how the IPB-MSA&SO_4 datasets can be used. Because the data product represents the concentration of freshly formed sulfur aerosol species and the ML model does not account for

atmospheric transport, users must interpret the datasets considering the air mass history. To better clarify the idea, we employed the independent MSA data measured during the Polarstern campaigns in the NA (Huang et al., 2017), which were not used in the training/validation or testing/evaluation of the ML models, and compared them with predicted MSA by GPR. In particular, the MSA by GPR was extracted along air mass BTs arriving at the hourly sites of the ship tracks and then averaged considering a 0-day (simultaneously), 1-day, 2-day and 3-day air mass history. The MSA measurements on Polarstern were performed in four scientific cruises including two spring seasons (April-May 2011 & April-May 2012) and two autumn seasons (October-November 2011 & October-November 2012). The ship tracks of the cruises from which the data were taken in the present study are shown in Fig. 7. It can be seen that the best match between GPR-simulated MSA and observed MSA occurred when 2-day air masses were considered. At 2-day air mass history, the slope reached ~~0.84~~0.78 and the correlation coefficient 0.81 (Fig. 7a-d). Again, as seen in Fig. 7f, GPR MSA is considerably more consistent with observations than CAMS, for which a significant difference with observations ($p < 0.05$) can be appreciated.

4.6 Spatial distributions ~~Monthly~~ of MSA and nss-$SO_4^=$ ~~SO_4~~ ~~distributions~~

In order to elucidate the geographical distributions of biogenic sulfur aerosol production across the NA domain, the IPB-MSA&SO_4 datasets in the 25 years (1998–2022) were averaged to obtain the climatic annual and monthly distributions of MSA and nss-$SO_4^=$ ~~SO_4~~ as illustrated in Fig. 8~~a~~ and Fig. 9~~8b, respectively~~. Across the NA domain, the annual average of MSA is 0.016 ± 0.007 μg m^{-3}, whereas the annual average of nss-$SO_4^=$ is 0.250 ±0.077 μg m^{-3} (Table S2). The annual spatial distributions of MSA and nss-$SO_4^=$ exhibit a latitudinal gradient over the majority of the NA area that increases from north to south, except below nearly 35° N, where it increases from west to east. Notwithstanding, the latitudinal gradients are much more evident than longitudinal variations. For instance, MSA grows at a rate of 0.0016 (R^2=0.93; p<0.05) μg m^{-3} per each 1° latitude towards the south and 0.00036 (R^2=0.53; p<0.05) μg m^{-3} per each 1° longitude eastward (it reaches its peak between 20° and 10° W). Furthermore, for each 1° southward, nss-$SO_4^=$ increases by 0.0212 (R^2=0.96; p<0.05) μg m^{-3}, whereas there are no significant changes in nss-$SO_4^=$ with longitude (R^2=0.01; p>0.05). The highest concentrations of both components (>90th percentile) are primarily found in the southeast of the domain (in front of the Moroccan coast and the Gibraltar strait). Minimum annual concentrations (<10th percentile) are found in northern areas of the domain, particularly in the Labrador Sea and near the shores of Greenland and Iceland.

The annual average MSA to nss-$SO_4^=$ (MSA:nss-$SO_4^=$) ratio is 0.053 ± 0.012 (Table S2), with a consistent latitudinal gradient increasing southward (rate of change = 0.0028 (R^2=0.93; p<0.05) per each 1° latitude). The lower MSA: nss-$SO_4^=$ values are found in the northwest of the domain, while the higher values are apparent in front of the African coast, the ratio is practically constant across the same latitudinal band. It is worth evidencing that the region with extremely high MSA concentrations and high MSA:nss-$SO_4^=$ (above the mean + three times the standard deviations) is linked to the Canary

upwelling system on the northwest African coast. The Canary Current system is one of the world's most productive regions of the ocean, known as eastern boundary upwelling systems (EBUSs) (Chavez and Messié, 2009; Carr, 2001). This may indicate a link between EBUSs and the potential formation of biogenic aerosol concentrations in the atmosphere. Previous research has shown how EBUSs changed in response to climate change (Bograd et al., 2023; Sydeman et al., 2014; Bonino et al., 2019), including the trend toward increased upwelling intensity (Wang et al., 2015; García-Reyes et al., 2015); however, little is known about the impact of EBUSs on marine biogenic emissions and the resultant aerosol fluxes. Future studies are needed to address these issues in order to better understand the role of EBUSs on the aerosol-climate systems.

Looking at ~~The~~ the monthly climatological maps (Fig. 9), it is revealed that MSA and nss-$SO_4^=$ ~~SO_4~~ display a gradual increase in their concentrations southward, clearly evident from October to March, resulting in a large difference between the northern and southern parts of the domain. On the~~In~~ contrary, during summer, the concentrations are more homogeneous over the domain (see latitudinal patterns in Fig. 9), still with a tendency to higher concentrations over the northeastern part. The seasonality of MSA and nss-$SO_4^=$ ~~SO_4~~ is evident: the increase for both compounds starts in April and peaks in June-July followed by a gradual decrease in September (Fig. ~~S8~~S9 and Table S2). The lowest MSA (nss-$SO_4^=$ ~~SO_4~~) concentration occurs in December at 0.006 ± 0.005 (0.155 ± 0.079) µg m^{-3} and the highest occurs in June at 0.029 ± 0.013 (0.364 ± 0.075) µg m^{-3} (~~Fig. 8a and Fig. 8b~~Table S2), consistent with the fact that winter and summer are typically the lowest and highest seasons for biological activity, respectively for the NA (Mansour et al., 2023a). The coefficient of variation (COV), defined as the ratio between the standard deviation and the mean value, expressed in percentage, at each grid point, is used to assess how much the MSA and nss-$SO_4^=$ vary around their mean value in each month; variability increases with higher COV. The maps (Fig. S10) confirm that the variability of sulfur aerosol species depends strongly on the season; MSA and nss-$SO_4^=$ are mostly stable (little variations) during the winter, whereas most variations occur between April (late spring) and June (early summer) and preferentially over the eastern part of the NA than the western.

The ~~ratio of MSA to SO4 (~~MSA:nss-$SO_4^=$ ~~SO_4)~~ also exhibits a seasonal pattern, with the lowest (highest) values observed during the winter (summer), as presented in Fig. ~~8c~~9c. July has the highest spatial average of the ratio of 0.077 ± 0.022 while the lowest of 0.032 ± 0.012 occurs in December (Table S2). Looking at the overall distributions, MSA:nss-$SO_4^=$ ~~SO_4~~ demonstrates a general ~~southern~~ southward increase, with the exception of summer months. In summer (mainly July and August), MSA:nss-$SO_4^=$ ~~SO_4~~ above 50°N has an opposite trend with respect to the one below 50°N. In detail, from North to South, we report a sharp increase in MSA:nss-$SO_4^=$ ~~SO_4~~, maximized around 50°N, followed by an abrupt decrease toward the equator. The possible explanation for the decline in MSA:nss-$SO_4^=$ ~~SO_4~~ below 50°N is that the reduction in MSA:nss-$SO_4^=$ ~~SO_4~~ ~~correlates to~~is related to an increase in AT caused by warmer air nearing the equator, in line with observations in the Pacific Ocean (Bates et al., 1992) and with the higher ratio observed in colder air masses (marine Polar and Arctic) with respect to warmer ones (marine Tropical) at MHD (Ovadnevaite et al., 2014). As a final remark, we report that the summertime low MSA:nss-$SO_4^=$ ~~SO_4~~ below 50°N is linked to a decrease in F_{DMS} in the same latitudinal zone (Mansour et al., 2023a). Owing to

the low DMS emissions, the different DMS oxidation patterns may be in competition (Barone et al., 1995); since MSA is formed preferentially through the pathway of OH addition at low temperatures (Shen et al., 2022), the production of MSA may be decreased relative to that of nss-$SO_4^=$ ~~SO_4~~ in the warm southern part of the domain, during summer, leading to the observed decrease in the MSA:nss-$SO_4^=$ ~~SO_4~~ ratio.

5. Data availability

The dataset includes daily MSA and nss-$SO_4^=$ ~~SO_4~~ concentrations at 0.25° × 0.25° spatial resolution over the North Atlantic Ocean from January 1998 to December 2022. The datasets are publicly available in NetCDF format as daily files on the Mendeley online repository at https://doi.org/10.17632/j8bzd5dvpx.1 (Mansour et al., 2023b).

6. Conclusions

Marine aerosol data can be obtained from in-situ coastal observatories or from shipborne measurements, however, punctual coast observations are limited under the point of view of the spatial representativity, while shipborne measurements suffer of limitations in terms of temporal coverage. Understanding the dynamics of marine-derived biogenic sulfur aerosols and their radiative effects, as well as carrying out relevant scientific studies, requires long-term, continuous and high-resolution (space and time-wise) datasets. To overcome the limitations of punctual measurements, we combined the in-situ observations of sulfur aerosol data at Mace Head and from NAAMES cruises, as dependent variables, and the sea-to-air DMS flux and ECMWF-ERA5 reanalysis meteorological datasets, as independent variables, to investigate the potential of machine learning techniques for the prediction of daily MSA and nss-$SO_4^=$ ~~SO_4~~ sea-level concentrations over the North Atlantic Ocean. We evaluated ~~four~~ five machine learning models (*i.e.,* SVM, DT, RE, GPR, and ANN), considering various sets of hyperparameter optimizations. Our findings demonstrated that the GPR model outperforms other approaches in simulating the concentrations of biogenic sulfur aerosols, capturing up to 86% and 72% of the observed variance in daily MSA and nss-$SO_4^=$ ~~SO_4~~, respectively. This makes the GPR an effective tool for obtaining trustworthy sea-level MSA and nss-$SO_4^=$ ~~SO_4~~ concentrations over the North Atlantic, which may also be successful in other oceanic regions or over the entire global ocean. The impact of the six independent predictors on the simulated MSA and nss-$SO_4^=$ ~~SO_4~~ is further evaluated using the GPR partial dependence analysis, which reveals that the relationships between them are multifaceted rather than linear or monotonically varying.

By the GPR machine learning method, we constructed a novel 0.25°×0.25° resolution daily gridded dataset of in-situ produced biogenic MSA and nss-$SO_4^=$ ~~SO_4~~ concentrations (named IPB-MSA&SO_4) covering the North Atlantic Ocean from 1998 to 2022. The dataset represents the sea-level concentrations of MSA and nss-$SO_4^=$ ~~SO_4~~ associated with in-situ production in the MBL, i.e., the concentration of sulfur aerosol species resulting, in each pixel, from the local biogenic emissions in

combination with local atmospheric conditions. Other inputs, such as terrestrial emissions or sinking of sulfur species produced in the free troposphere are not accounted for in the present dataset.

Comparison of the GPR-derived MSA with existing CAMS-EAC4 reanalysis product reveals that our high-resolution dataset accurately reproduces the spatial and temporal patterns of the biogenic sulfur aerosol concentration and has high consistency with the independent observations of the Polarstern cruises measurements in the Atlantic. The obtained IPB-MSA&SO_4 data were used to analyze the spatiotemporal variations of MSA, nss-$SO_4^=$~~SO_4~~, and the ratio between them (MSA:nss-$SO_4^=$~~SO_4~~). It was found that the monthly concentrations of MSA and nss-$SO_4^=$~~SO_4~~ across the NA are characterized by a significant southward increase in each month, with the exception of summertime when MSA and nss-$SO_4^=$~~SO_4~~ displayed more homogeneous spatial patterns with a tendency to higher concentrations over the northeastern part of the domain. The MSA:nss-$SO_4^=$~~SO_4~~ ratio exhibits a seasonal variation from winter (low) to summer (high) characterized by a sharp decline from the 50 °N parallel toward the equator mainly in July-August. In general, the atmospheric concentration of sulfur aerosol species tends to be more stable in winter, whereas wider variations are associated with late spring and early summertime and more with the eastern part of the domain than to the western one.

More ~~profound~~ in-depth analyses can be conducted based on the presented biogenic sulfur aerosol concentration dataset~~s~~, which could help further understanding of oceanic sulfur-aerosol-cloud interactions. For instance, we evidence that the Canary eastern upwelling system emerges from the dataset as a hotspot of high sea-level MSA concentration and high MSA:nss-$SO_4^=$ ratio; such a finding is worth further investigation and may shed light on the role of EBUSs in the production of biogenic marine aerosols and on its climate relevance.

Author contributions

KM and MR contributed to the conceptualization and design of the study. KM organized the datasets, constructed the models, analysed the data, and visualized the results. KM wrote the first draft of the manuscript under the supervision of MR. KM, MR, SD, DC, JO, LMR, MP, LP, SH, and CO contributed to the results investigation, manuscript revision, reading and editing, and approved the submitted version.

Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgements

We gratefully acknowledge the Copernicus climate change service (C3S) for the provision of ECMWF-ERA5 reanalysis meteorological data and the NOAA Air Resources Laboratory (ARL) for the provision of the HYSPLIT transport and dispersion model. University of Galway team acknowledges the support from Irish EPA (AC3 and AEROSOURCE, 2016-CCRP-MS-31) and the Department of Environment, Climate and Communications as well as MaREI, the SFI Research Centre for Energy, Climate, and Marine.

Financial support

The research was funded by the European Commission, H2020 Research Infrastructures, project FORCeS (grant no. 821205).

Figure 1: The study region of the North Atlantic Ocean (72° – 0° W, 20° – 66° N) with bathymetry presented in meters. The gridded bathymetric dataset was extracted from the General Bathymetric Chart of the Oceans (https://www.gebco.net), the GEBCO_2023 Grid. The ~~blue~~green-filled pentagram~~square~~ represents the Mace Head measuring station on the west coast of Ireland and the dark-red points are the sampling points that represent marine conditions in the NAAMES cruises track. The violet points represent the ship track during Polarstern campaigns.

Figure 2: The methodology's workflow. Predictors and response variables data preparation, the overall framework of generation and development of the trained models, including a schematic diagram of 5-fold cross-validation, models export and validation details, as well as post-processing analysis.

Figure 3: Comparison of predicted and observed MSA on the hourly (left panels) and daily (right panels) scales: (a) GPR, (b) EBT, (c) SVM, ~~and~~ (d) ANN, and (e) DT. The validation and test data subsets are used to compute the model's performance. R^2 and RMSE are computed in a logarithmic space, whereas MAE is computed on a normal scale.

Figure 4: Comparison of predicted and observed nss-$SO_4^{=}$~~SO_4~~ on the hourly (left panels) and daily (right panels) scales: (a) GPR, (b) EBT, (c) SVM, (d) ANN, and (e) DT~~, and (d) ANN~~. The validation and test data subsets are used to compute the model's performance. R^2 and RMSE are computed in a logarithmic space, whereas MAE is computed on a normal scale.

Figure 5: Partial dependence plots of MSA and nss-$SO_4^{=}$~~SO_4~~ as a function of the predictors revealed by the GPR model.

Figure 6: Comparison between observed MSA at MHD measuring site and both MSA predicted by GPR (a) and MSA extracted from CAMS reanalysis (b). (c) and (d): joint probability histograms between observed MSA and residual errors (observed–predicted); the black dashed lines represent the change of MSA residual errors in each bin. MAE is the mean absolute error, and the relative MAE has been calculated as the MAE divided by the range of observed MSA. (e) and (f): frequency distributions of the residual errors. (g): Seasonal box charts from different datasets. Each box chart displays the median (line inside of each box), the 1st and 3rd quartiles (bottom and top edges of each box), the minimum and maximum values that are not outliers (whiskers), and any outliers represented by '+' (computed as values that are more than 1.5 of the interquartile range away from the top or bottom of the box). Box charts whose notches (the shaded region around each median) do not overlap have different medians at the 95% confidence level.

Figure 7: (a) Scatter plots between observed MSA during the Polarstern campaigns (Huang et al., 2017) and predicted MSA by GPR, considering (a) 0-day, (b) 1-day, (c) 2-day and (d) 3-day air mass history. (f): Seasonal box charts from different datasets. The features displayed on each box chart are the same as those given in Fig. 6.

Figure 8: The annual averages of (a) MSA, (b) nss-$SO_4^=$, and (c) MSA:nss-$SO_4^=$ spatial distributions based on GPR at 0.25° × 0.25° resolution during the 1998–2022 period. The latitudinal (longitudinal) gradients of each component are displayed in the left (bottom) panels whereas shaded areas represent ± standard deviations. The black crosses evidence the extremely high concentrations, more than three times standard deviations plus the annual mean climatology.

Figure ~~8~~9: Monthly spatial distributions of (a) MSA ~~(μg m^{-3})~~, (b) nss-$SO_4^=$ ~~SO_4 (μg m^{-3})~~, and (c) MSA:nss-$SO_4^=$ ~~SO_4~~ based on GPR over 1998–2022 at 0.25° × 0.25° resolution. ~~The monthly (average ± spatial standard deviation) are shown in brackets above each panel.~~ (d) Monthly latitudinal distributions of each component while ~~MSA, SO_4, and the MSA:SO_4 based on GPR over 1998–2022. S~~ shaded areas represent ± standard deviations.

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

Table ~~1~~2: The Pearson's Coefficients between possible predictors at the selected marine air masses and the in-situ observed MSA and nss-$SO_4^=$ ~~SO_4~~ concentrations. The MSA, nss-$SO_4^=$ ~~SO_4~~ and F_{DMS} values are used in the log scale. All values are statistically significant at $p<0.05$. Bolds evidence the maximum during different days of air mass history.

Table ~~2~~3: Details of the number of hourly (MSA and nss-$SO_4^=$ ~~SO_4~~) data points corresponding to selected marine BTs. The threshold used for filtering outlier values, and the number of data points after filtering are given.

Table ~~3~~4: Multilinear regression of MSA and nss-$SO_4^=$ ~~SO_4~~ as a function of predictors. The MSA, nss-$SO_4^=$ and F_{DMS} values are used in the log scale. Each independent variable's contribution to R^2 is the decrease in total R^2 when that variable is eliminated. Individual R^2 contributions are normalized and added together to equal the overall R^2. According to the analysis

of variance (ANOVA) on the multilinear regression models, all predictors contribute statistically significantly ($p<0.05$) to the MSA and nss-$SO_4^{=}$ ~~SO_4~~ variance.

675

References

Allan, J. D., Jimenez, J. L., Williams, P. I., Alfarra, M. R., Bower, K. N., Jayne, J. T., Coe, H., and Worsnop, D. R.: Quantitative sampling using an Aerodyne aerosol mass spectrometer - 1. Techniques of data interpretation and error analysis, Journal of Geophysical Research-Atmospheres, 108, 10.1029/2002jd002358, 2003.

Asante-Okyere, S., Shen, C. B., Ziggah, Y. Y., Rulegeya, M. M., and Zhu, X. F.: Investigating the Predictive Performance of Gaussian Process Regression in Evaluating Reservoir Porosity and Permeability, Energies, 11, 10.3390/en11123261, 2018.

Barone, S. B., Turnipseed, A. A., and Ravishankara, A. R.: Role of adducts in the atmospheric oxidation of dimethyl sulfide, Faraday Discussions, 100, 39-54, 10.1039/fd9950000039, 1995.

Bates, T. S., Calhoun, J. A., and Quinn, P. K.: VARIATIONS IN THE METHANESULFONATE TO SULFATE MOLAR RATIO IN SUBMICROMETER MARINE AEROSOL-PARTICLES OVER THE SOUTH-PACIFIC OCEAN, Journal of Geophysical Research-Atmospheres, 97, 9859-9865, 10.1029/92jd00411, 1992.

Behrenfeld, M. J., Moore, R. H., Hostetler, C. A., Graff, J., Gaube, P., Russell, L. M., Chen, G., Doney, S. C., Giovannoni, S., Liu, H. Y., Proctor, C., Bolalios, L. M., Baetge, N., Davie-Martin, C., Westberry, T. K., Bates, T. S., Bell, T. G., Bidle, K. D., Boss, E. S., Brooks, S. D., Cairns, B., Carlson, C., Halsey, K., Harvey, E. L., Hu, C. M., Karp-Boss, L., Kleb, M., Menden-Deuer, S., Morison, F., Quinn, P. K., Scarino, A. J., Anderson, B., Chowdhary, J., Crosbie, E., Ferrare, R., Haire, J. W., Hu, Y. X., Janz, S., Redemann, J., Saltzman, E., Shook, M., Siegel, D. A., Wisthaler, A., Martine, M. Y., and Ziemba, L.: The North Atlantic Aerosol and Marine Ecosystem Study (NAAMES): Science Motive and Mission Overview, Frontiers in Marine Science, 6, 10.3389/fmars.2019.00122, 2019.

Bock, J., Michou, M., Nabat, P., Abe, M., Mulcahy, J. P., Olivie, D. J. L., Schwinger, J., Suntharalingam, P., Tjiputra, J., van Hulten, M., Watanabe, M., Yool, A., and Seferian, R.: Evaluation of ocean dimethylsulfide concentration and emission in CMIP6 models, Biogeosciences, 18, 3823-3860, 10.5194/bg-18-3823-2021, 2021.

Bograd, S. J., Jacox, M. G., Hazen, E. L., Lovecchio, E., Montes, I., Buil, M. P., Shannon, L. J., Sydeman, W. J., and Rykaczewski, R. R.: Climate Change Impacts on Eastern Boundary Upwelling Systems, Annual Review of Marine Science, 15, 303-328, 10.1146/annurev-marine-032122-021945, 2023.

Bonino, G., Di Lorenzo, E., Masina, S., and Iovino, D.: Interannual to decadal variability within and across the major Eastern Boundary Upwelling Systems, Scientific Reports, 9, 10.1038/s41598-019-56514-8, 2019.

Breiman, L.: Random forests, Machine Learning, 45, 5-32, 10.1023/a:1010933404324, 2001.

Buckley, M. W. and Marshall, J.: Observations, inferences, and mechanisms of the Atlantic Meridional Overturning Circulation: A review, Reviews of Geophysics, 54, 5-63, 10.1002/2015rg000493, 2016.

Canagaratna, M. R., Jayne, J. T., Jimenez, J. L., Allan, J. D., Alfarra, M. R., Zhang, Q., Onasch, T. B., Drewnick, F., Coe, H., Middlebrook, A., Delia, A., Williams, L. R., Trimborn, A. M., Northway, M. J., DeCarlo, P. F., Kolb, C. E., Davidovits, P., and Worsnop, D. R.: Chemical and microphysical characterization of ambient aerosols with the aerodyne aerosol mass spectrometer, Mass Spectrometry Reviews, 26, 185-222, 10.1002/mas.20115, 2007.

Carr, M. E.: Estimation of potential productivity in Eastern Boundary Currents using remote sensing, Deep-Sea Research Part Ii-Topical Studies in Oceanography, 49, 59-80, 2001.

Charlson, R. J., Lovelock, J. E., Andreae, M. O., and Warren, S. G.: OCEANIC PHYTOPLANKTON, ATMOSPHERIC SULFUR, CLOUD ALBEDO AND CLIMATE, Nature, 326, 655-661, 10.1038/326655a0, 1987.

Chavez, F. P. and Messié, M.: A comparison of Eastern Boundary Upwelling Ecosystems, Progress in Oceanography, 83, 80-96, 10.1016/j.pocean.2009.07.032, 2009.

DeCarlo, P. F., Kimmel, J. R., Trimborn, A., Northway, M. J., Jayne, J. T., Aiken, A. C., Gonin, M., Fuhrer, K., Horvath, T., Docherty, K. S., Worsnop, D. R., and Jimenez, J. L.: Field-deployable, high-resolution, time-of-flight aerosol mass spectrometer, Analytical Chemistry, 78, 8281-8289, 10.1021/ac061249n, 2006.

Etminan, M., Myhre, G., Highwood, E. J., and Shine, K. P.: Radiative forcing of carbon dioxide, methane, and nitrous oxide: A significant revision of the methane radiative forcing, Geophysical Research Letters, 43, 12614-12623, 10.1002/2016gl071930, 2016.

Fan, J. L., Yue, W. J., Wu, L. F., Zhang, F. C., Cai, H. J., Wang, X. K., Lu, X. H., and Xiang, Y. Z.: Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China, Agricultural and Forest Meteorology, 263, 225-241, 10.1016/j.agrformet.2018.08.019, 2018.

Fiddes, S. L., Woodhouse, M. T., Nicholls, Z., Lane, T. P., and Schofield, R.: Cloud, precipitation and radiation responses to large perturbations in global dimethyl sulfide, Atmospheric Chemistry and Physics, 18, 10177-10198, 10.5194/acp-18-10177-2018, 2018.

Fratantoni, D. M.: North Atlantic surface circulation during the 1990's observed with satellite-tracked drifters, Journal of Geophysical Research-Oceans, 106, 22067-22093, 10.1029/2000jc000730, 2001.

Friedman, J. H.: Greedy function approximation: A gradient boosting machine, Annals of Statistics, 29, 1189-1232, 10.1214/aos/1013203451, 2001.

Frossard, A. A., Russell, L. M., Massoli, P., Bates, T. S., and Quinn, P. K.: Side-by-Side Comparison of Four Techniques Explains the Apparent Differences in the Organic Composition of Generated and Ambient Marine Aerosol Particles, Aerosol Science and Technology, 48, V-X, 10.1080/02786826.2013.879979, 2014.

Fung, K. M., Heald, C. L., Kroll, J. H., Wang, S. Y., Jo, D. S., Gettelman, A., Lu, Z., Liu, X. H., Zaveri, R. A., Apel, E. C., Blake, D. R., Jimenez, J. L., Campuzano-Jost, P., Veres, P. R., Bates, T. S., Shilling, J. E., and Zawadowicz, M.: Exploring dimethyl sulfide (DMS) oxidation and implications for global aerosol radiative forcing, Atmospheric Chemistry and Physics, 22, 1549-1573, 10.5194/acp-22-1549-2022, 2022.

Fushiki, T.: Estimation of prediction error by using K-fold cross-validation, Statistics and Computing, 21, 137-146, 10.1007/s11222-009-9153-8, 2011.

Gali, M., Levasseur, M., Devred, E., Simo, R., and Babin, M.: Sea-surface dimethylsulfide (DMS) concentration from satellite data at global and regional scales, Biogeosciences, 15, 3497-3519, 10.5194/bg-15-3497-2018, 2018.

García-Reyes, M., Sydeman, W. J., Schoeman, D. S., Rykaczewski, R. R., Black, B. A., Smit, A. J., and Bograd, S. J.: Under Pressure: Climate Change, Upwelling, and Eastern Boundary Upwelling Ecosystems, Frontiers in Marine Science, 2, 10.3389/fmars.2015.00109, 2015.

Goddijn-Murphy, L., Woolf, D. K., and Marandino, C.: Space-based retrievals of air-sea gas transfer velocities using altimeters: Calibration for dimethyl sulfide, Journal of Geophysical Research-Oceans, 117, 10.1029/2011jc007535, 2012.

Gondwe, M., Krol, M., Gieskes, W., Klaassen, W., and de Baar, H.: The contribution of ocean-leaving DMS to the global atmospheric burdens of DMS, MSA, SO2, and NSS SO4=, Global Biogeochemical Cycles, 17, 10.1029/2002gb001937, 2003.

Grigas, T., Ovadnevaite, J., Ceburnis, D., Moran, E., McGovern, F. M., Jennings, S. G., and O'Dowd, C.: Sophisticated Clean Air Strategies Required to Mitigate Against Particulate Organic Pollution, Scientific Reports, 7, 10.1038/srep44737, 2017.

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., and Graler, B.: Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables, Peerj, 6, 10.7717/peerj.5518, 2018.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horanyi, A., Munoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Holm, E., Janiskova, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thepaut, J. N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999-2049, 10.1002/qj.3803, 2020.

Hodshire, A. L., Campuzano-Jost, P., Kodros, J. K., Croft, B., Nault, B. A., Schroder, J. C., Jimenez, J. L., and Pierce, J. R.: The potential role of methanesulfonic acid (MSA) in aerosol formation and growth and the associated radiative forcings, Atmospheric Chemistry and Physics, 19, 3137-3160, 10.5194/acp-19-3137-2019, 2019.

Huang, S., Poulain, L., van Pinxteren, D., van Pinxteren, M., Wu, Z. J., Herrmann, H., and Wiedensohler, A.: Latitudinal and Seasonal Distribution of Particulate MSA over the Atlantic using a Validated Quantification Method with HR-ToF-AMS, Environmental Science & Technology, 51, 418-426, 10.1021/acs.est.6b03186, 2017.

Hulswar, S., Simo, R., Gali, M., Bell, T. G., Lana, A., Inamdar, S., Halloran, P. R., Manville, G., and Mahajan, A. S.: Third revision of the global surface seawater dimethyl sulfide climatology (DMS-Rev3), Earth System Science Data, 14, 2963-2987, 10.5194/essd-14-2963-2022, 2022.

Inness, A., Ades, M., Agusti-Panareda, A., Barre, J., Benedictow, A., Blechschmidt, A. M., Dominguez, J. J., Engelen, R., Eskes, H., Flemming, J., Huijnen, V., Jones, L., Kipling, Z., Massart, S., Parrington, M., Pench, V. H., Razinger, M., Remy, S., Schulz, M., and Suttie, M.: The CAMS reanalysis of atmospheric composition, Atmospheric Chemistry and Physics, 19, 3515-3556, 10.5194/acp-19-3515-2019, 2019.

Jimenez, J. L., Jayne, J. T., Shi, Q., Kolb, C. E., Worsnop, D. R., Yourshaw, I., Seinfeld, J. H., Flagan, R. C., Zhang, X. F., Smith, K. A., Morris, J. W., and Davidovits, P.: Ambient aerosol sampling using the Aerodyne Aerosol Mass Spectrometer, Journal of Geophysical Research-Atmospheres, 108, 10.1029/2001jd001213, 2003.

Kalogirou, S. A.: Artificial neural networks in renewable energy systems applications: a review, Renewable & Sustainable Energy Reviews, 5, 373-401, 10.1016/s1364-0321(01)00006-5, 2001.

Katsman, C. A., Spall, M. A., and Pickart, R. S.: Boundary current eddies and their role in the restratification of the Labrador Sea, Journal of Physical Oceanography, 34, 1967-1983, 10.1175/1520-0485(2004)034<1967:bceatr>2.0.co;2, 2004.

Kim, A. H., Yum, S. S., Lee, H., Chang, D. Y., and Shim, S.: Polar Cooling Effect Due to Increase of Phytoplankton and Dimethyl-Sulfide Emission, Atmosphere, 9, 10.3390/atmos9100384, 2018.

Kotsiantis, S. B.: Decision trees: a recent overview, Artificial Intelligence Review, 39, 261-283, 10.1007/s10462-011-9272-4, 2013.

Langmann, B., Scannell, C., and O'Dowd, C.: New Directions: Organic matter contribution to marine aerosols and cloud condensation nuclei, Atmospheric Environment, 42, 7821-7822, 10.1016/j.atmosenv.2008.09.002, 2008.

Lins, I. D., Araujo, M., Moura, M. D., Silva, M. A., and Droguett, E. L.: Prediction of sea surface temperature in the tropical Atlantic by support vector machines, Computational Statistics & Data Analysis, 61, 187-198, 10.1016/j.csda.2012.12.003, 2013.

Mahajan, A. S., Fadnavis, S., Thomas, M. A., Pozzoli, L., Gupta, S., Royer, S. J., Saiz-Lopez, A., and Simo, R.: Quantifying the impacts of an updated global dimethyl sulfide climatology on cloud microphysics and aerosol radiative forcing, Journal of Geophysical Research-Atmospheres, 120, 2524-2536, 10.1002/2014jd022687, 2015.

Mahmood, R., von Salzen, K., Norman, A. L., Gali, M., and Levasseur, M.: Sensitivity of Arctic sulfate aerosol and clouds to changes in future surface seawater dimethylsulfide concentrations, Atmospheric Chemistry and Physics, 19, 6419-6435, 10.5194/acp-19-6419-2019, 2019.

Mansour, K., Decesari, S., Ceburnis, D., Ovadnevaite, J., and 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, 871, 10.1016/j.scitotenv.2023.162123, 2023a.

Mansour, K., Decesari, S., Ceburnis, D., Ovadnevaite, J., Russell, L., Paglione, M., O'Dowd, C., and Rinaldi, M.: IPB-MSA&SO4: In-situ Produced Biogenic Methanesulfonic Acid and Sulfate over the North Atlantic (V1) [dataset], 10.17632/j8bzd5dvpx.1, 2023b.

Mansour, K., Rinaldi, M., Preissler, J., Decesari, S., Ovadnevaite, J., Ceburnis, D., Paglione, M., Facchini, M. C., and O'Dowd, C.: Phytoplankton Impact on Marine Cloud Microphysical Properties Over the Northeast Atlantic Ocean, Journal of Geophysical Research-Atmospheres, 127, 10.1029/2021jd036355, 2022.

Mansour, K., Decesari, S., Bellacicco, M., Marullo, S., Santoleri, R., Bonasoni, P., Facchini, M. C., Ovadnevaite, J., Ceburnis, D., O'Dowd, C., and Rinaldi, M.: Particulate methanesulfonic acid over the central Mediterranean Sea: Source region identification and relationship with phytoplankton activity, Atmospheric Research, 237, 10.1016/j.atmosres.2019.104837, 2020a.

Mansour, K., Decesari, S., Facchini, M. C., Belosi, F., Paglione, M., Sandrini, S., Bellacicco, M., Marullo, S., Santoleri, R., Ovadnevaite, J., Ceburnis, D., O'Dowd, C., Roberts, G., Sanchez, K., and Rinaldi, M.: Linking Marine Biological Activity to Aerosol Chemical Composition and Cloud-Relevant Properties Over the North Atlantic Ocean, Journal of Geophysical Research-Atmospheres, 125, 10.1029/2019jd032246, 2020b.

Marzocchi, A., Hirschi, J. J. M., Holliday, N. P., Cunningham, S. A., Blaker, A. T., and Coward, A. C.: The North Atlantic subpolar circulation in an eddy-resolving global ocean model, Journal of Marine Systems, 142, 126-143, 10.1016/j.jmarsys.2014.10.007, 2015.

McNabb, B. J. and Tortell, P. D.: Improved prediction of dimethyl sulfide (DMS) distributions in the northeast subarctic Pacific using machine-learning algorithms, Biogeosciences, 19, 1705-1721, 10.5194/bg-19-1705-2022, 2022.

Mendes-Moreira, J., Soares, C., Jorge, A. M., and De Sousa, J. F.: Ensemble Approaches for Regression: A Survey, Acm Computing Surveys, 45, 10.1145/2379776.2379786, 2012.

O'Dowd, C., Ceburnis, D., Ovadnevaite, J., Vaishya, A., Rinaldi, M., and Facchini, M. C.: Do anthropogenic, continental or coastal aerosol sources impact on a marine aerosol signature at Mace Head?, Atmospheric Chemistry and Physics, 14, 10687-10704, 10.5194/acp-14-10687-2014, 2014.

O'Dowd, C., Ceburnis, D., Ovadnevaite, J., Bialek, J., Stengel, D. B., Zacharias, M., Nitschke, U., Connan, S., Rinaldi, M., Fuzzi, S., Decesari, S., Facchini, M. C., Marullo, S., Santoleri, R., Dell'Anno, A., Corinaldesi, C., Tangherlini, M., and Danovaro, R.: Connecting marine productivity to sea-spray via nanoscale biological processes: Phytoplankton Dance or Death Disco?, Scientific Reports, 5, 10.1038/srep14883, 2015.

O'Dowd, C. D., Facchini, M. C., Cavalli, F., Ceburnis, D., Mircea, M., Decesari, S., Fuzzi, S., Yoon, Y. J., and Putaud, J. P.: Biogenically driven organic contribution to marine aerosol, Nature, 431, 676-680, 10.1038/nature02959, 2004.

Ovadnevaite, J., Ceburnis, D., Leinert, S., Dall'Osto, M., Canagaratna, M., O'Doherty, S., Berresheim, H., and O'Dowd, C.: Submicron NE Atlantic marine aerosol chemical composition and abundance: Seasonal trends and air mass categorization, Journal of Geophysical Research-Atmospheres, 119, 11850-11863, 10.1002/2013jd021330, 2014.

Quinlan, J. R.: Induction of decision trees, Machine Learning, 1, 81-106, 10.1007/BF00116251, 1986.

Quinn, P. K. and Bates, T. S.: The case against climate regulation via oceanic phytoplankton sulphur emissions, Nature, 480, 51-56, 10.1038/nature10580, 2011.

Rhein, M., Kieke, D., Huttl-Kabus, S., Roessler, A., Mertens, C., Meissner, R., Klein, B., Boning, C. W., and Yashayaev, I.: Deep water formation, the subpolar gyre, and the meridional overturning circulation in the subpolar North Atlantic, Deep-Sea Research Part Ii-Topical Studies in Oceanography, 58, 1819-1832, 10.1016/j.dsr2.2010.10.061, 2011.

Riccobono, F., Schobesberger, S., Scott, C. E., Dommen, J., Ortega, I. K., Rondo, L., Almeida, J., Amorim, A., Bianchi, F., Breitenlechner, M., David, A., Downard, A., Dunne, E. M., Duplissy, J., Ehrhart, S., Flagan, R. C., Franchin, A., Hansel, A., Junninen, H., Kajos, M., Keskinen, H., Kupc, A., Kurten, A., Kvashin, A. N., Laaksonen, A., Lehtipalo, K., Makhmutov, V., Mathot, S., Nieminen, T., Onnela, A., Petaja, T., Praplan, A. P., Santos, F. D., Schallhart, S., Seinfeld, J. H., Sipila, M., Spracklen, D. V., Stozhkov, Y., Stratmann, F., Tome, A., Tsagkogeorgas, G., Vaattovaara, P., Viisanen, Y., Vrtala, A., Wagner, P. E., Weingartner, E., Wex, H., Wimmer, D., Carslaw, K. S., Curtius, J., Donahue, N. M., Kirkby, J., Kulmala, M., Worsnop, D. R., and Baltensperger, U.: Oxidation Products of Biogenic Emissions Contribute to Nucleation of Atmospheric Particles, Science, 344, 717-721, 10.1126/science.1243527, 2014.

Rinaldi, M., Decesari, S., Finessi, E., Giulianelli, L., Carbone, C., Fuzzi, S., O'Dowd, C. D., Ceburnis, D., and Facchini, M. C.: Primary and Secondary Organic Marine Aerosol and Oceanic Biological Activity: Recent Results and New Perspectives for Future Studies, Advances in Meteorology, 10.1155/2010/310682, 2010.

Rinaldi, M., Hiranuma, N., Santachiara, G., Mazzola, M., Mansour, K., Paglione, M., Rodriguez, C. A., Traversi, R., Becagli, S., Cappelletti, D., and Belosi, F.: Ice-nucleating particle concentration measurements from Ny-Alesund during the Arctic spring-summer in 2018, Atmospheric Chemistry and Physics, 21, 14725-14748, 10.5194/acp-21-14725-2021, 2021.

Rinaldi, M., Facchini, M. C., Decesari, S., Carbone, C., Finessi, E., Mircea, M., Fuzzi, S., Ceburnis, D., Ehn, M., Kulmala, M., de Leeuw, G., and O'Dowd, C. D.: On the representativeness of coastal aerosol studies to open ocean studies: Mace Head - a case study, Atmospheric Chemistry and Physics, 9, 9635-9646, 10.5194/acp-9-9635-2009, 2009.

Rodriguez, J. D., Perez, A., and Lozano, J. A.: Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation, Ieee Transactions on Pattern Analysis and Machine Intelligence, 32, 569-575, 10.1109/tpami.2009.187, 2010.

Rolph, G., Stein, A., and Stunder, B.: Real-time Environmental Applications and Display sYstem: READY, Environmental Modelling & Software, 95, 210-228, 10.1016/j.envsoft.2017.06.025, 2017.

Royer, S. J., Mahajan, A. S., Gali, M., Saltzman, E., and Simo, R.: Small-scale variability patterns of DMS and phytoplankton in surface waters of the tropical and subtropical Atlantic, Indian, and Pacific Oceans, Geophysical Research Letters, 42, 475-483, 10.1002/2014gl062543, 2015.

Sachindra, D. A., Ahmed, K., Rashid, M. M., Shahid, S., and Perera, B. J. C.: Statistical downscaling of precipitation using machine learning techniques, Atmospheric Research, 212, 240-258, 10.1016/j.atmosres.2018.05.022, 2018.

Saliba, G., Chen, C. L., Lewis, S., Russell, L. M., Quinn, P. K., Bates, T. S., Bell, T. G., Lawler, M. J., Saltzman, E. S., Sanchez, K. J., Moore, R., Shook, M., Rivellini, L. H., Lee, A., Baetge, N., Carlson, C. A., and Behrenfeld, M. J.: Seasonal Differences and Variability of Concentrations, Chemical Composition, and Cloud Condensation Nuclei of Marine Aerosol Over the North Atlantic, Journal of Geophysical Research-Atmospheres, 125, 10.1029/2020jd033145, 2020.

Sanchez, K. J., Chen, C. L., Russell, L. M., Betha, R., Liu, J., Price, D. J., Massoli, P., Ziemba, L. D., Crosbie, E. C., Moore, R. H., Muller, M., Schiller, S. A., Wisthaler, A., Lee, A. K. Y., Quinn, P. K., Bates, T. S., Porter, J., Bell, T. G., Saltzman, E. S., Vaillancourt, R. D., and Behrenfeld, M. J.: Substantial Seasonal Contribution of Observed Biogenic Sulfate Particles to Cloud Condensation Nuclei, Scientific Reports, 8, 10.1038/s41598-018-21590-9, 2018.

Sarker, I. H., Kayes, A. S. M., and Watters, P.: Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage, Journal of Big Data, 6, 10.1186/s40537-019-0219-y, 2019.

Shabani, S., Samadianfard, S., Sattari, M. T., Mosavi, A., Shamshirband, S., Kmet, T., and Varkonyi-Koczy, A. R.: Modeling Pan Evaporation Using Gaussian Process Regression K-Nearest Neighbors Random Forest and Support Vector Machines; Comparative Analysis, Atmosphere, 11, 10.3390/atmos11010066, 2020.

Shen, J. L., Scholz, W., He, X. C., Zhou, P. T., Marie, G., Wang, M. Y., Marten, R., Surdu, M., Rorup, B., Baalbaki, R., Amorim, A., Ataei, F., Bell, D. M., Bertozzi, B., Brasseur, Z., Caudillo, L., Chen, D. X., Chu, B. W., Dada, L., Duplissy, J., Finkenzeller, H., Granzin, M., Guida, R., Heinritzi, M., Hofbauer, V., Iyer, S., Kemppainen, D., Kong, W. M., Krechmer, J. E., Kurten, A., Lamkaddam, H., Lee, C. P., Lopez, B., Mahfouz, N. G. A., Manninen, H. E., Massabo, D., Mauldin, R. L.,

Mentler, B., Muller, T., Pfeifer, J., Philippov, M., Piedehierro, A. A., Roldin, P., Schobesberger, S., Simon, M., Stolzenburg, D., Tham, Y. J., Tome, A., Umo, N. S., Wang, D. Y., Wang, Y. H., Weber, S. K., Welti, A., de Jonge, R. W., Wu, Y. S., Zauner-Wieczorek, M., Zust, F., Baltensperger, U., Curtius, J., Flagan, R. C., Hansel, A., Mohler, O., Petaja, T., Volkamer, R., Kulmala, M., Lehtipalo, K., Rissanen, M., Kirkby, J., El-Haddad, I., Bianchi, F., Sipila, M., Donahue, N. M., and Worsnop, D. R.: High Gas-Phase Methanesulfonic Acid Production in the OH-Initiated Oxidation of Dimethyl Sulfide at Low Temperatures, Environmental Science & Technology, 56, 13931-13944, 10.1021/acs.est.2c05154, 2022.

Shrestha, N. K. and Shukla, S.: Support vector machine based modeling of evapotranspiration using hydro-climatic variables in a sub-tropical environment, Agricultural and Forest Meteorology, 200, 172-184, 10.1016/j.agrformet.2014.09.025, 2015.

Simo, R.: Production of atmospheric sulfur by oceanic plankton: biogeochemical, ecological and evolutionary links, Trends in Ecology & Evolution, 16, 287-294, 10.1016/s0169-5347(01)02152-8, 2001.

Stein, A. F., Draxler, R. R., Rolph, G. D., Stunder, B. J. B., Cohen, M. D., and Ngan, F.: NOAA'S HYSPLIT ATMOSPHERIC TRANSPORT AND DISPERSION MODELING SYSTEM, Bulletin of the American Meteorological Society, 96, 2059-2077, 10.1175/bams-d-14-00110.1, 2015.

Sydeman, W. J., García-Reyes, M., Schoeman, D. S., Rykaczewski, R. R., Thompson, S. A., Black, B. A., and Bograd, S. J.: Climate change and wind intensification in coastal upwelling ecosystems, Science, 345, 77-80, 10.1126/science.1251635, 2014.

Thomas, M. A., Suntharalingam, P., Pozzoli, L., Rast, S., Devasthale, A., Kloster, S., Feichter, J., and Lenton, T. M.: Quantification of DMS aerosol-cloud-climate interactions using the ECHAM5-HAMMOZ model in a current climate scenario, Atmospheric Chemistry and Physics, 10, 7425-7438, 10.5194/acp-10-7425-2010, 2010.

Vapnik, V. N.: The Nature of Statistical Learning Theory, 2, Springer New York, NY, 10.1007/978-1-4757-3264-1, 2013.

Verrelst, J., Rivera, J. P., Gitelson, A., Delegido, J., Moreno, J., and Camps-Valls, G.: Spectral band selection for vegetation properties retrieval using Gaussian processes regression, International Journal of Applied Earth Observation and Geoinformation, 52, 554-567, 10.1016/j.jag.2016.07.016, 2016.

von Glasow, R. and Crutzen, P. J.: Model study of multiphase DMS oxidation with a focus on halogens, Atmospheric Chemistry and Physics, 4, 589-608, 10.5194/acp-4-589-2004, 2004.

Wang, D. W., Gouhier, T. C., Menge, B. A., and Ganguly, A. R.: Intensification and spatial homogenization of coastal upwelling under climate change, Nature, 518, 390-394, 10.1038/nature14235, 2015.

Williams, C. K. I. and Rasmussen, C. E.: Gaussian processes for regression, Advances in Neural Information Processing Systems 8: Proceedings of the 1995 Conference, 8, 514-520, 1996.

Woodhouse, M. T., Carslaw, K. S., Mann, G. W., Vallina, S. M., Vogt, M., Halloran, P. R., and Boucher, O.: Low sensitivity of cloud condensation nuclei to changes in the sea-air flux of dimethyl-sulphide, Atmospheric Chemistry and Physics, 10, 7545-7559, 10.5194/acp-10-7545-2010, 2010.

Zhou, S. Q., Chen, Y., Wang, F. H., Bao, Y., Ding, X. P., and Xu, Z. J.: 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, 10.1021/acs.est.3c04382, 2023.

Zhou, S. Q., Chen, Y., Paytan, A., Li, H. W., Wang, F. H., Zhu, Y. C., Yang, T. J., Zhang, Y., and Zhang, R. F.: Non-Marine Sources Contribute to Aerosol Methanesulfonate Over Coastal Seas, Journal of Geophysical Research-Atmospheres, 126, 10.1029/2021jd034960, 2021.

Zhu, L., Nenes, A., Wine, P. H., and Nicovich, J. M.: Effects of aqueous organosulfur chemistry on particulate methanesulfonate to non-sea salt sulfate ratios in the marine atmosphere, Journal of Geophysical Research-Atmospheres, 111, 10.1029/2005jd006326, 2006.

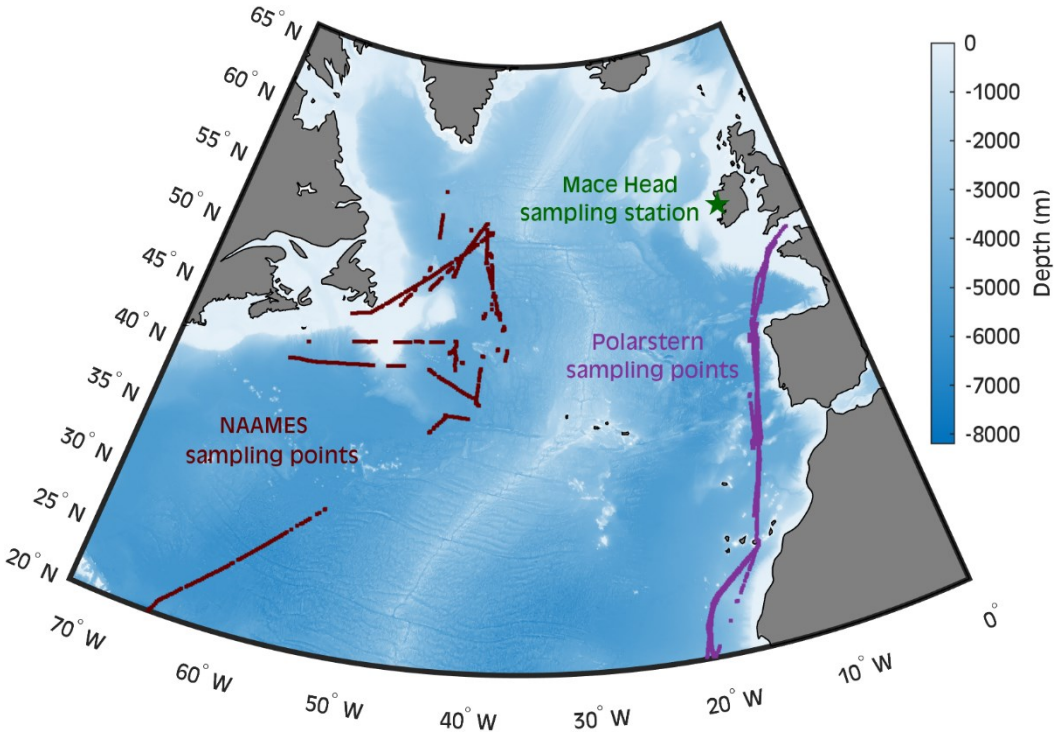

65° N
60° N
55° N
50° N
45° N
40° N
35° N
30° N
25° N
20° N
70° W
60° W
50° W
40° W
30° W
20° W
10° W
0°
Mace Head
sampling station
Polarstern
sampling points
NAAMES
sampling points
0
-1000
-2000
-3000
-4000
-5000
-6000
-7000
-8000
Depth (m)

Fig. 1

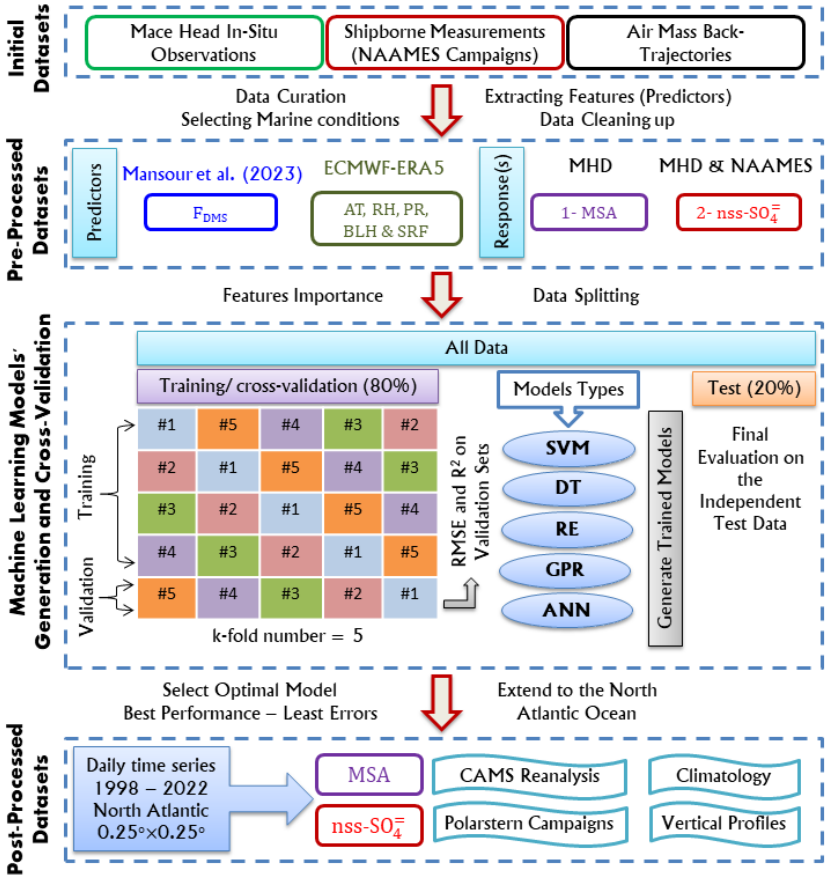

Initial Datasets
Mace Head In-Situ Observations
Shipborne Measurements (NAAMES Campaigns)
Air Mass Back-Trajectories
Data Curation
Selecting Marine conditions
Extracting Features (Predictors)
Data Cleaning up
Pre-Processed Datasets
Predictors
Mansour et al. (2023)
ECMWF-ERA5
F_{DMS}
AT, RH, PR, BLH & SRF
Response(s)
MHD
MHD & NAAMES
1- MSA
2- nss-$SO_4^=$
Features Importance
Data Splitting
Machine Learning Models' Generation and Cross-Validation
All Data
Training/ cross-validation (80%)
Test (20%)
Training
Validation
k-fold number = 5
RMSE and R^2 on Validation Sets
Models Types
SVM
DT
RE
GPR
ANN
Generate Trained Models
Final Evaluation on the Independent Test Data
Select Optimal Model
Best Performance – Least Errors
Extend to the North Atlantic Ocean
Post-Processed Datasets
Daily time series
1998 – 2022
North Atlantic
0.25°×0.25°
MSA
nss-$SO_4^=$
CAMS Reanalysis
Polarstern Campaigns
Climatology
Vertical Profiles

Fig. 2

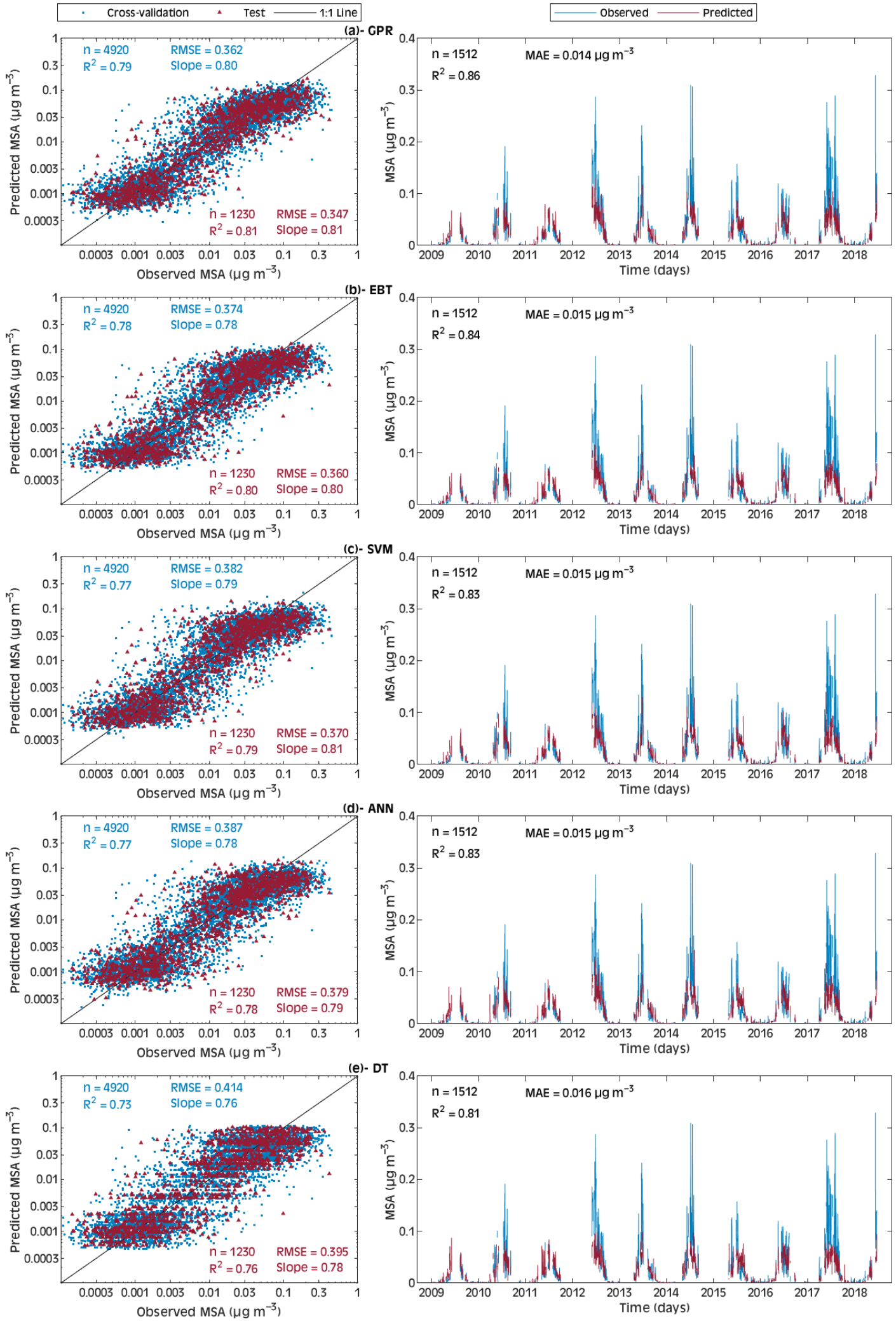

Cross-validation
Test
1:1 Line
Observed
Predicted
(a)- GPR
n = 4920
RMSE = 0.362
R2 = 0.79
Slope = 0.80
n = 1230
RMSE = 0.347
R2 = 0.81
Slope = 0.81
n = 1512
MAE = 0.014 µg m−3
R2 = 0.86
(b)- EBT
n = 4920
RMSE = 0.374
R2 = 0.78
Slope = 0.78
n = 1230
RMSE = 0.360
R2 = 0.80
Slope = 0.80
n = 1512
MAE = 0.015 µg m−3
R2 = 0.84
(c)- SVM
n = 4920
RMSE = 0.382
R2 = 0.77
Slope = 0.79
n = 1230
RMSE = 0.370
R2 = 0.79
Slope = 0.81
n = 1512
MAE = 0.015 µg m−3
R2 = 0.83
(d)- ANN
n = 4920
RMSE = 0.387
R2 = 0.77
Slope = 0.78
n = 1230
RMSE = 0.379
R2 = 0.78
Slope = 0.79
n = 1512
MAE = 0.015 µg m−3
R2 = 0.83
(e)- DT
n = 4920
RMSE = 0.414
R2 = 0.73
Slope = 0.76
n = 1230
RMSE = 0.395
R2 = 0.76
Slope = 0.78
n = 1512
MAE = 0.016 µg m−3
R2 = 0.81
Predicted MSA (µg m−3)
Observed MSA (µg m−3)
MSA (µg m−3)
Time (days)
0.0003
0.001
0.003
0.01
0.03
0.1
0.3
1
0
0.1
0.2
0.3
0.4
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018

Fig. 3

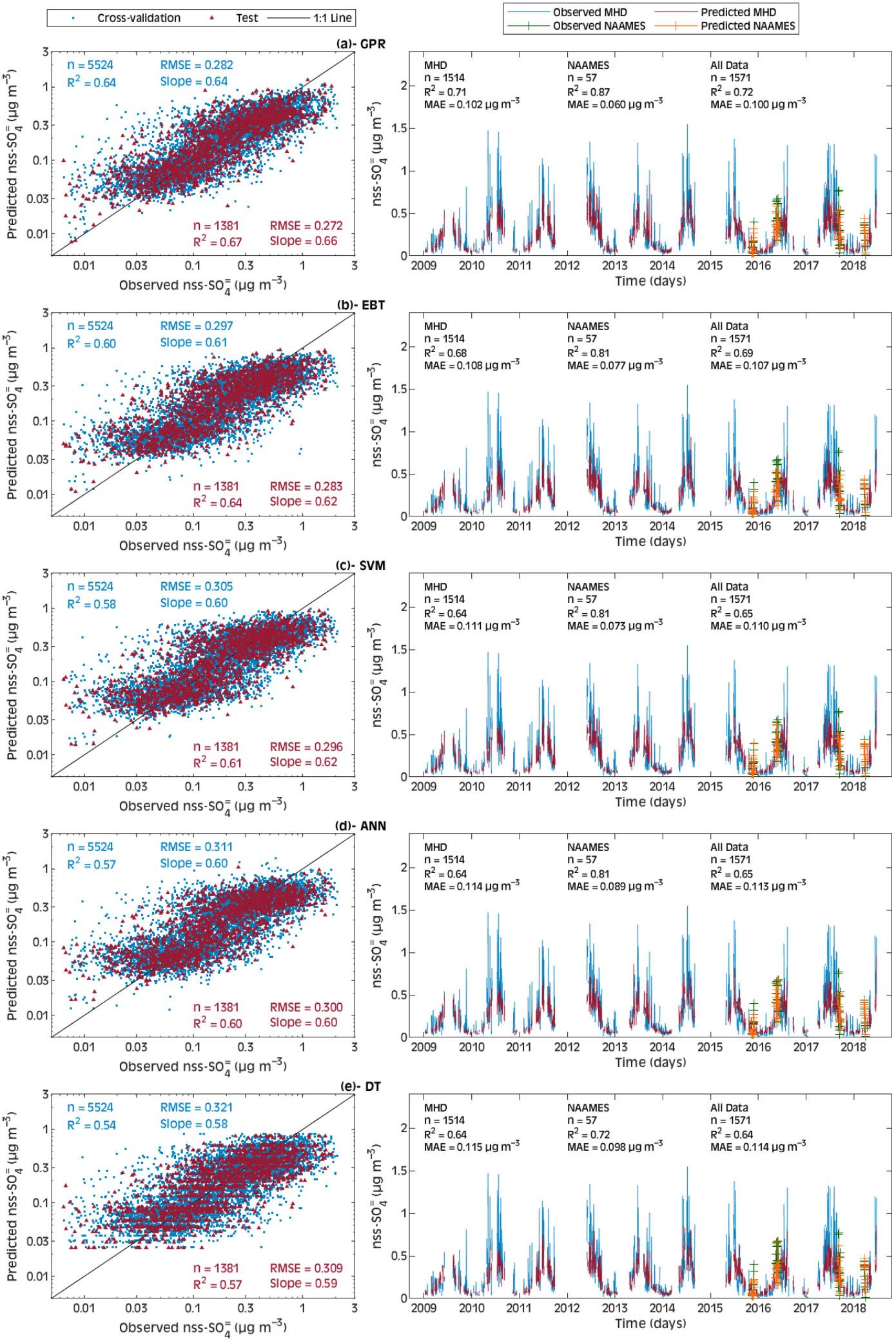

Cross-validation
Test
1:1 Line
Observed MHD
Predicted MHD
Observed NAAMES
Predicted NAAMES
(a)- GPR
n = 5524
R2 = 0.64
RMSE = 0.282
Slope = 0.64
n = 1381
R2 = 0.67
RMSE = 0.272
Slope = 0.66
MHD
n = 1514
R2 = 0.71
MAE = 0.102 μg m−3
NAAMES
n = 57
R2 = 0.87
MAE = 0.060 μg m−3
All Data
n = 1571
R2 = 0.72
MAE = 0.100 μg m−3
(b)- EBT
n = 5524
R2 = 0.60
RMSE = 0.297
Slope = 0.61
n = 1381
R2 = 0.64
RMSE = 0.283
Slope = 0.62
MHD
n = 1514
R2 = 0.68
MAE = 0.108 μg m−3
NAAMES
n = 57
R2 = 0.81
MAE = 0.077 μg m−3
All Data
n = 1571
R2 = 0.69
MAE = 0.107 μg m−3
(c)- SVM
n = 5524
R2 = 0.58
RMSE = 0.305
Slope = 0.60
n = 1381
R2 = 0.61
RMSE = 0.296
Slope = 0.62
MHD
n = 1514
R2 = 0.64
MAE = 0.111 μg m−3
NAAMES
n = 57
R2 = 0.81
MAE = 0.073 μg m−3
All Data
n = 1571
R2 = 0.65
MAE = 0.110 μg m−3
(d)- ANN
n = 5524
R2 = 0.57
RMSE = 0.311
Slope = 0.60
n = 1381
R2 = 0.60
RMSE = 0.300
Slope = 0.60
MHD
n = 1514
R2 = 0.64
MAE = 0.114 μg m−3
NAAMES
n = 57
R2 = 0.81
MAE = 0.089 μg m−3
All Data
n = 1571
R2 = 0.65
MAE = 0.113 μg m−3
(e)- DT
n = 5524
R2 = 0.54
RMSE = 0.321
Slope = 0.58
n = 1381
R2 = 0.57
RMSE = 0.309
Slope = 0.59
MHD
n = 1514
R2 = 0.64
MAE = 0.115 μg m−3
NAAMES
n = 57
R2 = 0.72
MAE = 0.098 μg m−3
All Data
n = 1571
R2 = 0.64
MAE = 0.114 μg m−3
Predicted nss-SO4= (μg m−3)
Observed nss-SO4= (μg m−3)
nss-SO4= (μg m−3)
Time (days)
0.01
0.03
0.1
0.3
1
3
0
0.5
1.5
2
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018

Fig. 4

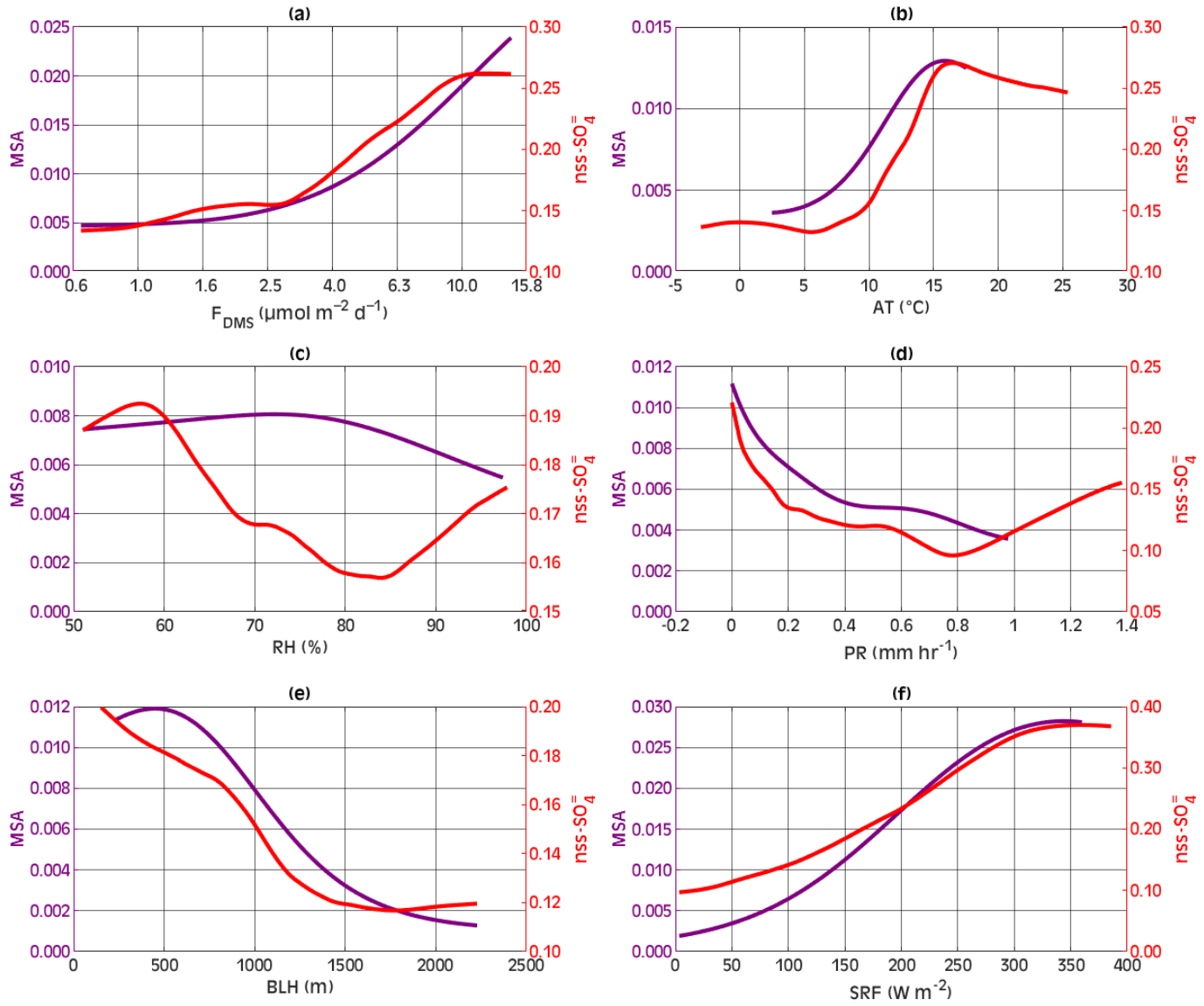

(a)
MSA
nss-SO4=
0.025
0.020
0.015
0.010
0.005
0.000
0.30
0.25
0.20
0.15
0.10
0.6
1.0
1.6
2.5
4.0
6.3
10.0
15.8
FDMS (μmol m−2 d−1)
(b)
MSA
nss-SO4=
0.015
0.010
0.005
0.000
0.30
0.25
0.20
0.15
0.10
-5
0
5
10
15
20
25
30
AT (°C)
(c)
MSA
nss-SO4=
0.010
0.008
0.006
0.004
0.002
0.000
0.20
0.19
0.18
0.17
0.16
0.15
50
60
70
80
90
100
RH (%)
(d)
MSA
nss-SO4=
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0.25
0.20
0.15
0.10
0.05
-0.2
0
0.2
0.4
0.6
0.8
1
1.2
1.4
PR (mm hr-1)
(e)
MSA
nss-SO4=
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0.20
0.18
0.16
0.14
0.12
0.10
0
500
1000
1500
2000
2500
BLH (m)
(f)
MSA
nss-SO4=
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0.40
0.30
0.20
0.10
0.00
0
50
100
150
200
250
300
350
400
SRF (W m-2)

Fig. 5

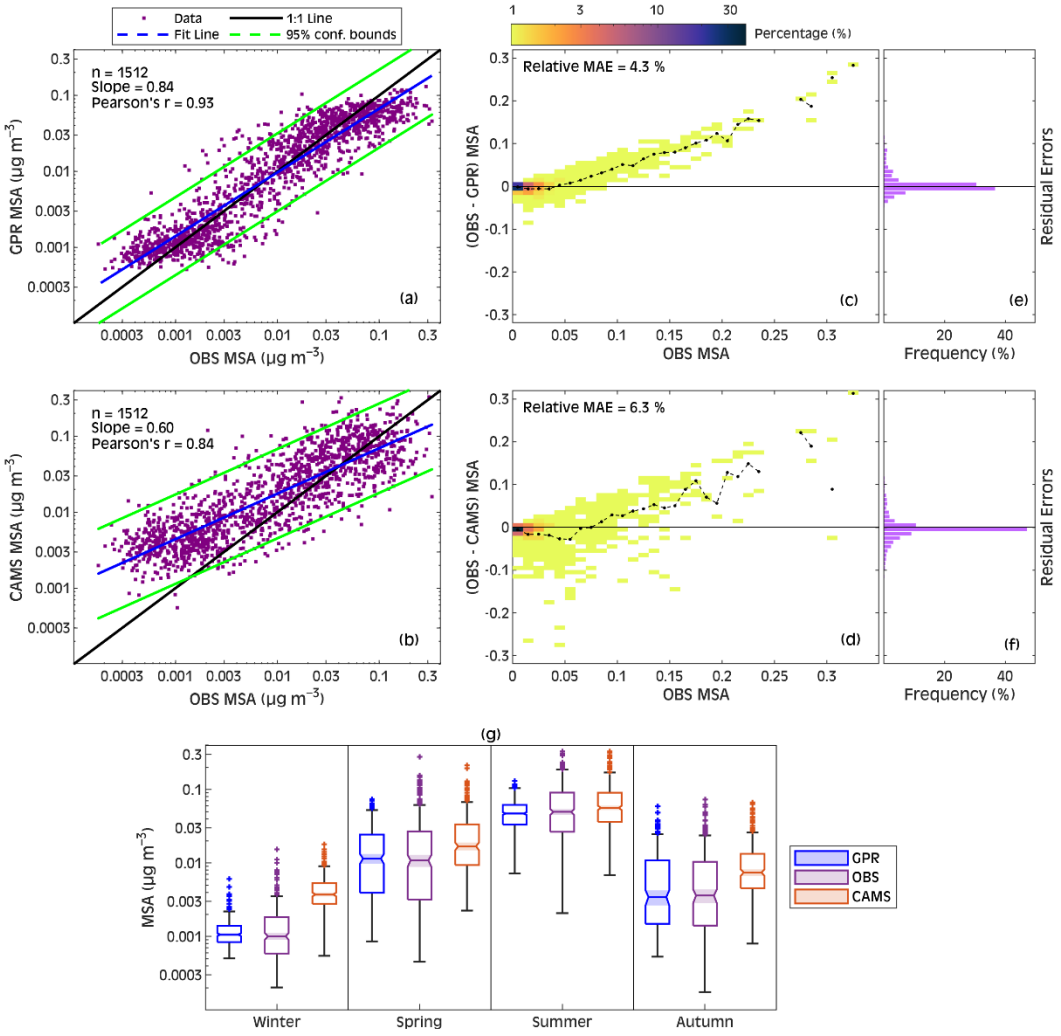
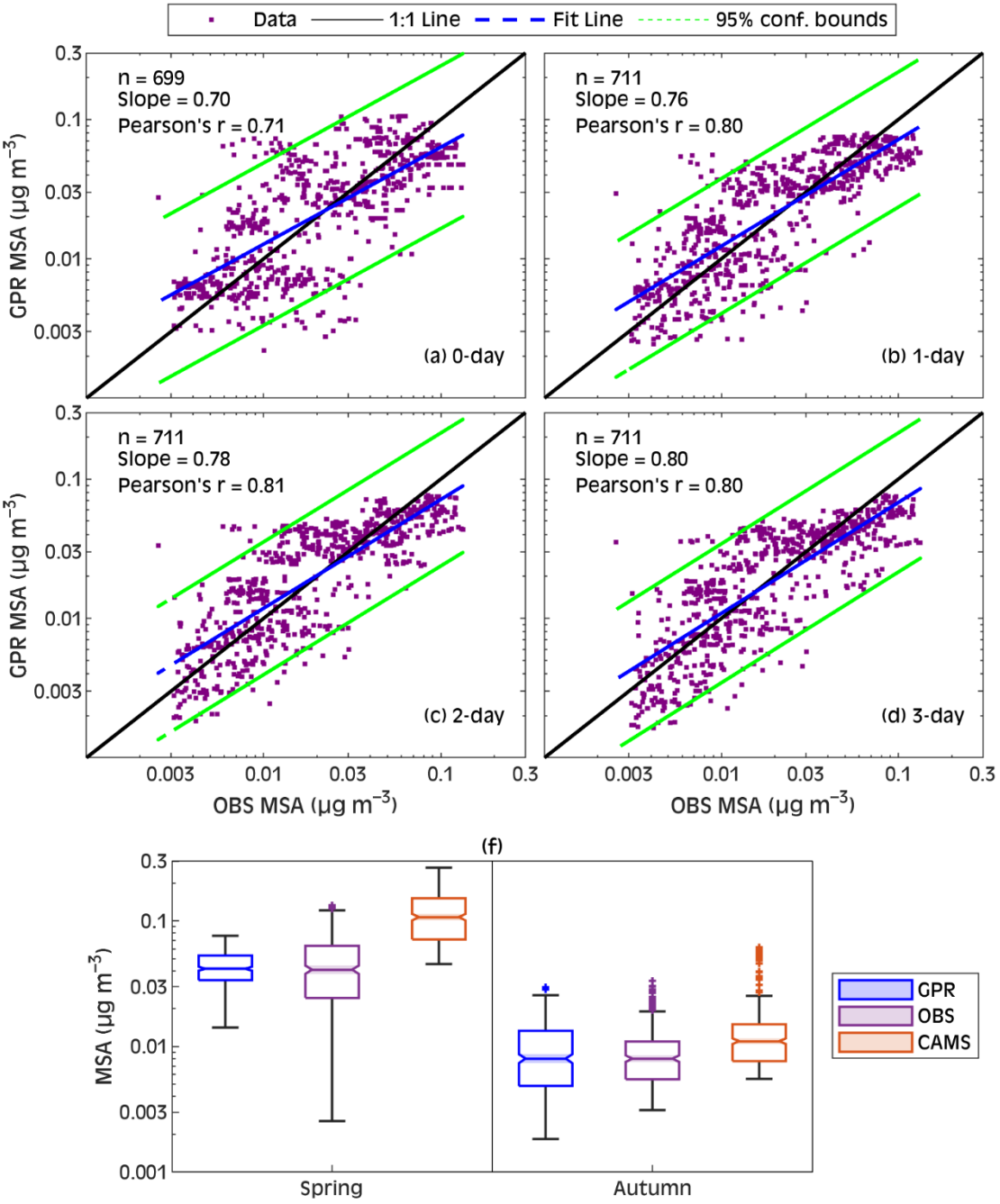

Data
1:1 Line
Fit Line
95% conf. bounds
n = 1512
Slope = 0.84
Pearson's r = 0.93
GPR MSA (μg m−3)
OBS MSA (μg m−3)
(a)
1
3
10
30
Percentage (%)
Relative MAE = 4.3 %
(OBS - GPR) MSA
OBS MSA
(c)
Residual Errors
Frequency (%)
(e)
n = 1512
Slope = 0.60
Pearson's r = 0.84
CAMS MSA (μg m−3)
OBS MSA (μg m−3)
(b)
Relative MAE = 6.3 %
(OBS - CAMS) MSA
OBS MSA
(d)
Residual Errors
Frequency (%)
(f)
(g)
MSA (μg m−3)
Winter
Spring
Summer
Autumn
GPR
OBS
CAMS

Fig. 6


Data
1:1 Line
Fit Line
95% conf. bounds
n = 699
Slope = 0.70
Pearson's r = 0.71
(a) 0-day
n = 711
Slope = 0.76
Pearson's r = 0.80
(b) 1-day
n = 711
Slope = 0.78
Pearson's r = 0.81
(c) 2-day
n = 711
Slope = 0.80
Pearson's r = 0.80
(d) 3-day
GPR MSA (μg m−3)
OBS MSA (μg m−3)
0.3
0.1
0.03
0.01
0.003

(f)
MSA (μg m−3)
0.3
0.1
0.03
0.01
0.003
0.001
Spring
Autumn
GPR
OBS
CAMS

Fig. 7

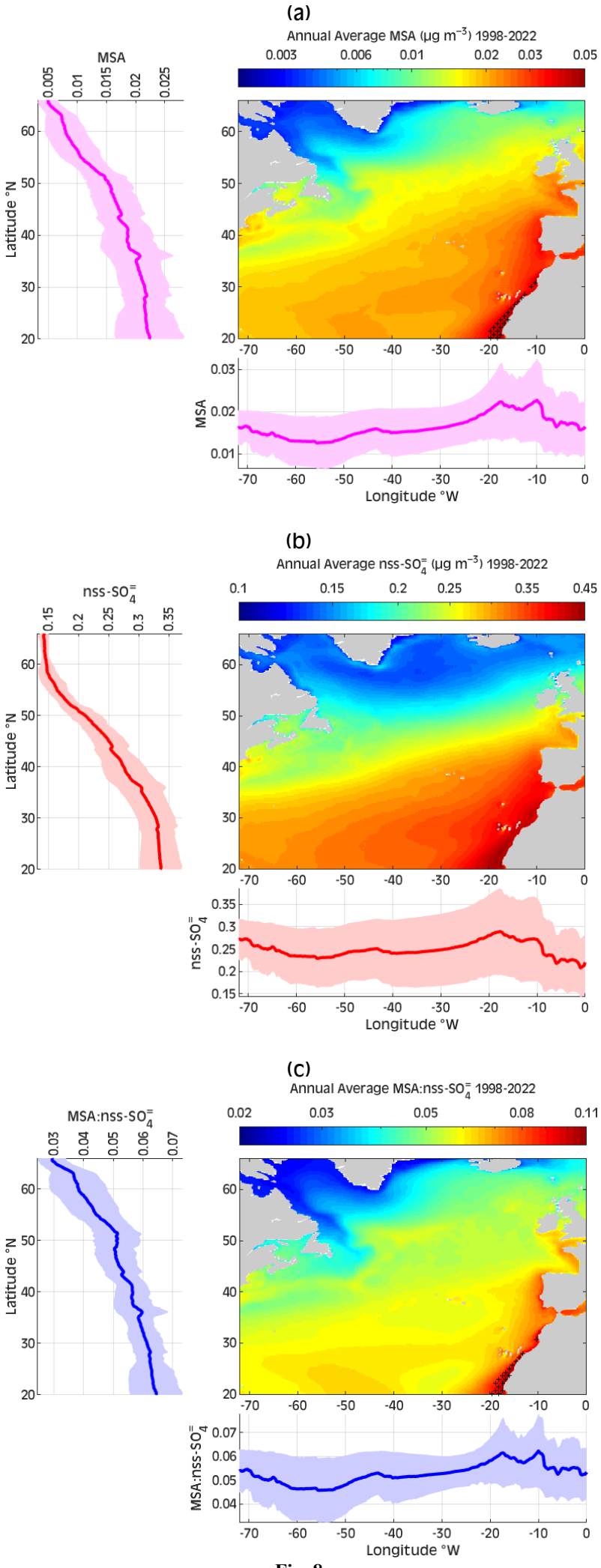

(a)
Annual Average MSA (μg m^{-3}) 1998-2022
MSA
Latitude °N
Longitude °W
(b)
Annual Average nss-SO$_4^=$ (μg m^{-3}) 1998-2022
nss-SO$_4^=$
(c)
Annual Average MSA:nss-SO$_4^=$ 1998-2022
MSA:nss-SO$_4^=$

Fig. 8

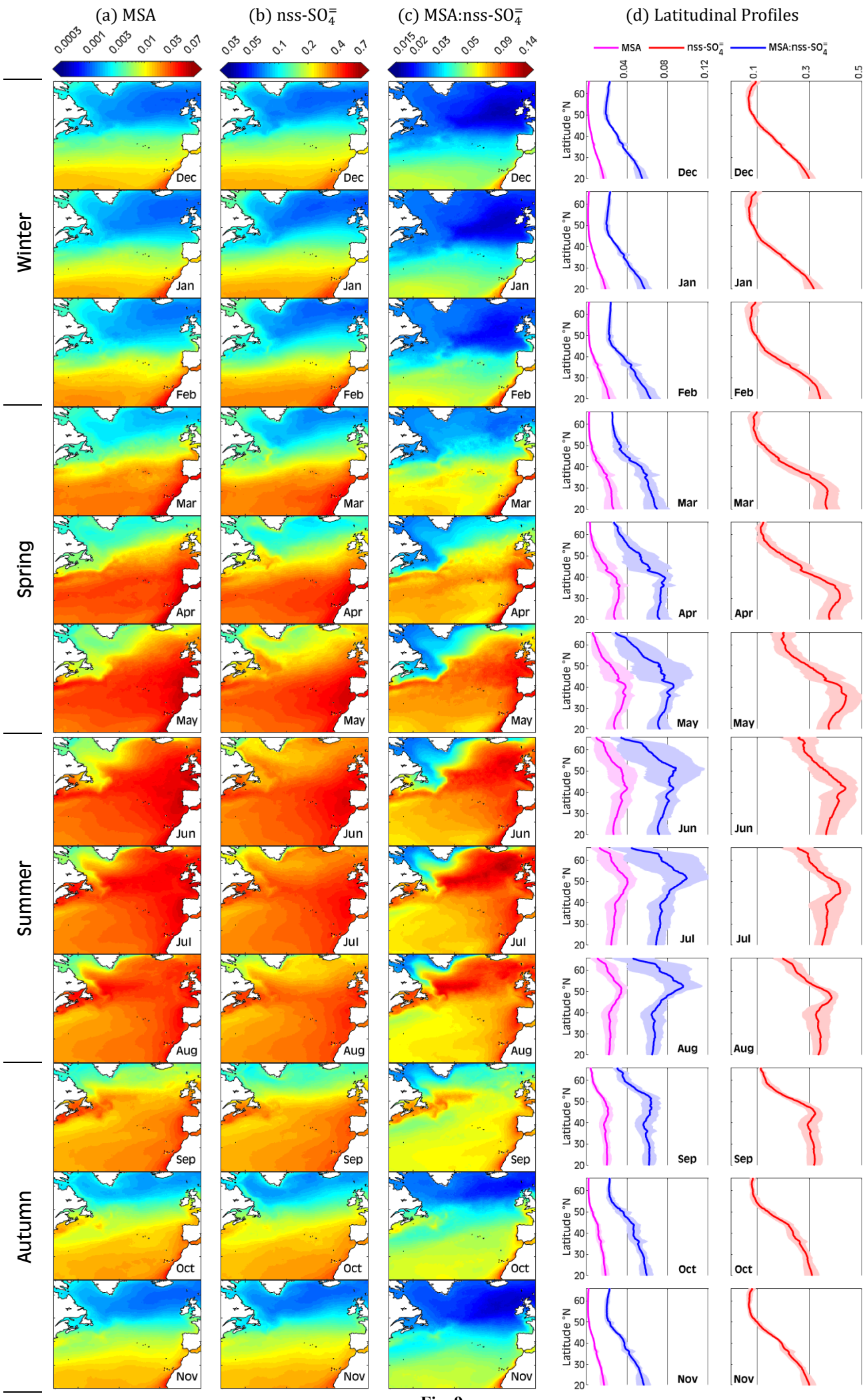

(a) MSA
(b) nss-SO$_4^=$
(c) MSA:nss-SO$_4^=$
(d) Latitudinal Profiles
MSA
nss-SO$_4^=$
MSA:nss-SO$_4^=$
Winter
Spring
Summer
Autumn
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Latitude °N

Fig. 9

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 1^{st} layer size = 100
	Bi-layered	Number of fully connected layers = 2 1^{st} layer size = 10 2^{nd} layer size = 10
	Tri-layered	Number of fully connected layers = 3 1^{st} layer size = 10 2^{nd} layer size = 10 3^{rd} layer size = 10

Table 1

Predictors	BT length (days)	MSA	nss-$SO_4^=$	
		MHD	MHD	NAAMES
F_{DMS}	0	0.27	0.24	0.04*
	1	0.64	0.53	0.24
	2	0.66	0.54	0.38
	3	**0.69**	**0.55**	**0.47**
AT	0	**0.65**	**0.61**	0.17
	1	0.57	0.56	0.29
	2	0.53	0.53	0.35
	3	0.53	0.51	**0.37**
RH	0	0.15	0.15	0.27
	1	0.33	0.27	0.22
	2	0.39	0.31	0.24
	3	**0.44**	**0.33**	**0.28**
PR	0	-0.18	-0.12	-0.09
	1	-0.27	-0.26	-0.27
	2	-0.33	-0.31	**-0.34**
	3	**-0.35**	**-0.33**	-0.32
BLH	0	-0.41	-0.32	-0.32
	1	-0.53	-0.45	-0.34
	2	-0.58	-0.49	**-0.36**
	3	**-0.60**	**-0.49**	-0.35
SRF	0	0.32	0.23	0.14
	1	0.73	0.61	0.53
	2	0.77	0.65	0.62
	3	**0.78**	**0.67**	**0.63**

Table 2

Response	No. of hourly data points	Lower threshold according to 0.1 percentile	Upper threshold according to 99.9 percentile	No. of data lost due to filtering	No. of hourly data points after cleanup
MSA	Source: MHD $n = 6162$	0.0001 $\mu g\ m^{-3}$	0.45 $\mu g\ m^{-3}$	12	6150
nss-$SO_4^{=}$	Source: MHD $n = 6260$	0.006 $\mu g\ m^{-3}$	2.116 $\mu g\ m^{-3}$	12	6905
	Source: NAAMES $n = 660$	0.007 $\mu g\ m^{-3}$	1.107 $\mu g\ m^{-3}$	3	

Table 3

	Total explained variance by R^2	Normalized Contribution to R^2 (%)					
		AT	RH	PR	BLH	SRF	F_{DMS}
MSA	74.36%	6.86	0.47	2.82	8.66	42.77	12.97
nss-$SO_4^=$	53.39%	11.64	0.55	5.07	3.63	25.83	6.66

Table 4