Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Machine learning for prediction of daily sea surface dimethylsulfide concentration and emission flux over the North Atlantic Ocean (1998–2021)

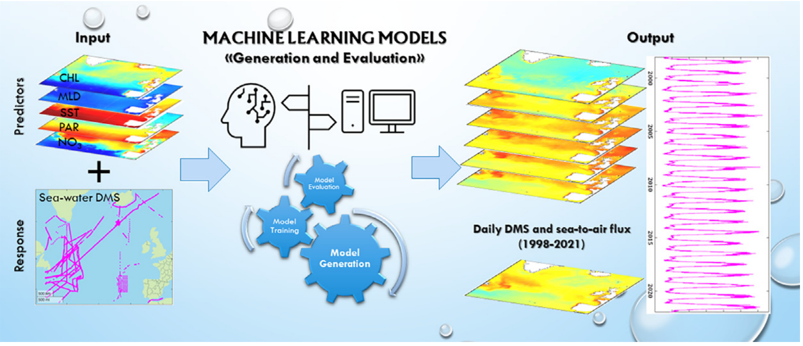
Karam Mansour[a,b,*], Stefano Decesari[a], Darius Ceburnis[c], Jurgita Ovadnevaite[c], Matteo Rinaldi[a]

[a] Italian National Research Council, Institute of Atmospheric Sciences and Climate (CNR-ISAC), Bologna 40129, Italy
[b] Oceanography Department, Faculty of Science, Alexandria University, Alexandria 21500, Egypt
[c] School of Physics, Ryan Institute's Centre for Climate and Air Pollution Studies, National University of Ireland Galway, Galway, Ireland

HIGHLIGHTS

- Dimethylsulfide (DMS) flux estimation requires a precise calculation of seawater DMS.
- Seawater DMS were calculated using machine learning models.
- Gaussian process can explain up to 71 % of DMS variance, outperforming other models.
- Phytoplankton biomass and ocean mixed layer depth control the DMS regional patterns.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Hai Guo

Keywords:
DMS
Sea-to-air flux
Gaussian process regression
K-means++
Biogenic sulfur emissions

ABSTRACT

As the most ubiquitous natural source of sulfur in the atmosphere, dimethylsulfide (DMS) promotes aerosol formation in marine environments, impacting cloud radiative forcing and precipitation, eventually influencing regional and global climate. In this study, we propose a machine learning predictive algorithm based on Gaussian process regression (GPR) to model the distribution of daily DMS concentrations in the North Atlantic waters over 24 years (1998–2021) at 0.25° × 0.25° spatial resolution. The model was built using DMS observations from cruises, combined with satellite-derived oceanographic data and Copernicus-modelled data. Further comparison was made with the previously employed machine learning methods (i.e., artificial neural network and random forest regression) and the existing empirical DMS algorithms. The proposed GPR outperforms the other methods for predicting DMS, displaying the highest coefficient of determination (R^2) value of 0.71 and the least root mean square error (RMSE) of 0.21. Notably, DMS regional patterns are associated with the spatial distribution of phytoplankton biomass and the thickness of the ocean mixed layer, displaying high DMS concentrations above 50°N from June to August. The amplitude, onset, and duration of the DMS annual cycle vary significantly across different regions, as revealed by the k-means++ clustering. Based on the GPR model output, the sea-to-air flux in the North Atlantic from March to September is estimated to be 3.04 Tg S, roughly 44 % lower than the estimates based on extrapolations of in-situ data. The present study demonstrates the effectiveness of a novel method for estimating seawater DMS surface concentration at unprecedented space and time resolutions. As a result, we are able to capture high-frequency spatial and temporal patterns in DMS variability. Better predictions of DMS concentration and derived sea-to-air flux will improve the modeling of biogenic sulfur aerosol concentrations in the atmosphere and reduce aerosol-cloud interaction uncertainties in climate models.

* Corresponding author at: Italian National Research Council, Institute of Atmospheric Sciences and Climate (CNR-ISAC), Bologna 40129, Italy.
E-mail address: k.mansour@isac.cnr.it (K. Mansour).

http://dx.doi.org/10.1016/j.scitotenv.2023.162123
Received 29 September 2022; Received in revised form 13 January 2023; Accepted 5 February 2023
Available online 10 February 2023

1. Introduction

Dimethylsulfide (DMS) is a volatile biogenic gas produced in seawater by marine phytoplankton and microbial metabolism, such as the decomposition of the algal metabolite dimethylsulfoniopropionate (DMSP) (Kettle et al., 1999). Once emitted into the atmosphere, DMS is oxidized yielding various sulfur products, the most important of which are methanesulfonic acid (MSA) and non-sea-salt sulfate (Charlson et al., 1987; Facchini et al., 2008; Mansour et al., 2020a). The sea-to-air global DMS flux (F_{DMS}) was estimated to be 28.1 Tg S per year (Lana et al., 2011), making it the largest biological source of sulfur aerosol in the atmospheric marine boundary layer. Such aerosols act as cloud condensation nuclei (CCN), potentially influencing Earth's albedo and climate (Charlson et al., 1987). Accurate parametrization of oceanic DMS is required for precise estimation of their flux, which is a key factor in understanding the Earth's climate feedback system. A variety of biotic and abiotic factors influence the oceanic DMS cycle (Mansour et al., 2020a), demonstrating its complexity and nonlinearity. The DMSP (precursor of DMS) is released into the dissolved oceanic pool through different mechanisms, including phytoplankton grazing (Wolfe and Steinke, 1996), viral lysis (Hill et al., 1998), stressed/ senescent cells (Laroche et al., 1999; Zhuang et al., 2011), and nutrient availability (Zindler et al., 2014). Physical variables such as salinity (Dickson and Kirst, 1987), temperature, and UV radiation (Toole and Siegel, 2004; Vallina and Simo, 2007) may also increase DMSP in marine algal cells by inducing stress (Sunda et al., 2002), which may regulate DMS release in seawater. Moreover, strong to storm wind speeds can alter DMS sea-to-air emission flux rates, resulting in significant depletions in surface water DMS concentrations (Royer et al., 2016); additionally, windy regions generally have deeply mixed waters where DMS cannot accumulate as it does in stratified waters.

The North Atlantic (NA) Ocean displays widespread seasonal phytoplankton blooms proxied by chlorophyll-*a* concentration (CHL), with distinct spatial dynamics (Friedland et al., 2016; Lacour et al., 2015). Observations revealed that elevated surface ocean DMS concentrations were generally linked to enhanced phytoplankton biomass (Bell et al., 2021) and exhibited marked seasonality over the NA Ocean (Lana et al., 2011). The biogenic emissions at the surface ocean including DMS-derived sulfate significantly contribute to the NA submicron marine aerosol burden (O'Dowd et al., 2004). The spatio-temporal correlations based on satellite ocean color data showed that the NA phytoplankton activity impacts sulfate aerosol chemical composition, aerosol number concentration, CCN (Mansour et al., 2020b) and ultimately cloud properties (Mansour et al., 2022). Precisely, variations in cloud droplet number concentrations caused by marine biota contribute to the enhanced albedo and are comparable in magnitude to those caused by anthropogenic inputs (Mansour et al., 2022). A summary of previously reported campaigns for seawater DMS measurements in the NA can be found in Table S1. In general, DMS concentrations have been found to vary both spatially and temporally throughout the studied domain. For instance, the highest DMS concentration was reported in a centrally located area of the East Atlantic Ocean between 56.0°N and 63.4°N and between 17.0°W and 22.4°W, with an average (median) of 8.93 (8.21) μmol m^{-3}, during June–July 1998. The North Atlantic Aerosols and Marine Ecosystems Study (NAAMES), a research project that aims to comprehend the relationships between ecosystems, aerosols, and clouds (Behrenfeld et al., 2019), conducted the most recent cruises for measuring surface ocean DMS concentrations. In the NW Atlantic over a 4-year period (2015–2018), four shipboard field campaigns were carried out as part of NAAMES. The cruises concluded that the variability in calculated sea-to-air F_{DMS} is seasonally dependent on the interplay of seawater DMS and wind speed variations, and inevitably on the choice of seawater DMS algorithm (Bell et al., 2021). The lowest calculated F_{DMS} of all NAAMES cruises occurred during NAAMES1, from 5th November to 2nd December 2015 (mean flux = 4.4 μmol m^{-2} d^{-1}). The highest average F_{DMS} was equal to 13.7 μmol m^{-2} d^{-1} during NAAMES4 (20th March–13th April 2018) as a result of a combined effect of higher wind speeds, compared to the other NAAMES cruises, and elevated seawater DMS levels in the latter half of the cruise. The bottom line, due to the scarcity and sparseness of observations, is that documenting the relationships between DMS concentration and the overlying aerosol/cloud properties in the highly dynamic NA environment remains a challenge. This study contributes to improving the prediction of seawater DMS spatial and temporal variability, explaining the complexity of the DMS cycle, and proposing the use of innovative machine learning predictive algorithms rather than simple- and multi- linear regressions.

Previous empirical approaches utilizing linear and multilinear regressions were developed to model DMS distributions (Gali et al., 2018; Simo and Dachs, 2002; Vallina and Simo, 2007). The approaches estimated DMS concentrations using several variables, including ratios of CHL to mixed layer depth (MLD) (Simo and Dachs, 2002), solar radiation dose (SRD) (Vallina and Simo, 2007), photosynthetically active radiation (PAR), and satellite-based DMSP concentrations (Gali et al., 2018). For more information on the empirical algorithms, see Section 2.4. The aforementioned methods have performed well in explaining DMS seasonal variations qualitatively (Wang et al., 2020), but their predictive power was generally reduced at regional scales (Herr et al., 2019), failing to accurately resolve the smaller-scale oceanic features (McNabb and Tortell, 2022; Royer et al., 2015). Identifying reliable high time-resolution data of DMS concentration is critical because variations in DMS dynamics frequently occur at the timescales of meteorological forcing, i.e., days to weeks (Royer et al., 2016).

Machine learning methods particularly artificial neural networks (ANN) and random forest regression (RFR) have been applied to derive monthly seawater DMS concentrations and make a prediction for the nonlinear oceanic systems. For example, the ANN was used to derive monthly climatology of global ocean DMS distributions (Wang et al., 2020) at 1° × 1° spatial resolution, yielding reasonable predictive skills ($R^2 = 0.66$) when compared to the raw data in the global database. However, the finer spatial patterns (e.g., mesoscale [roughly 20–200 km] and sub mesoscale [roughly 1–20 km]) and shorter temporal scale variability of oceanographic features driving DMS were not represented (Bell et al., 2021), due to the coarse space and time resolution. In the NA ocean, the area of interest of the present study, Bell et al. (2021) concluded that the ANN models consistently underestimated the in situ DMS concentrations, possibly because they were developed using climatological input parameters (Wang et al., 2020), which inevitably smooth out the episodic extremes in the predictor variables. Furthermore, two machine learning methods (RFR and ANN), were applied in the northeast subarctic Pacific (McNabb and Tortell, 2022) at a higher spatial resolution of 0.25° × 0.25°. The models captured up to 62 % of observed monthly seawater DMS variability and demonstrated notable regional patterns that are associated with mesoscale oceanographic variability which otherwise would be hidden using coarser spatial resolutions.

The present study aims at predicting the DMS concentrations in the NA surface waters at unprecedented spatial (0.25° × 0.25°) and temporal (daily) resolutions. Furthermore, the study proposes a new machine learning approach which is Gaussian process regression (GPR) that has never been applied to describe DMS distributions. GPR was chosen as the best-performing model for reconstructing the DMS observations after testing the most common machine learning regression models, including regression trees, support vector machines, regression ensembles, and artificial neural networks. Gaussian process, a powerful tool of machine learning algorithms, is a non-parametric kernel-based Bayesian probabilistic approach for solving regression problems (Williams and Rasmussen, 1996). It is a stochastic process that produces good results in terms of developing a calibration model for both linear and non-linear datasets; moreover, it can provide uncertainty measurements of the predictions. GPR works well on small datasets and the feature ranking is based on the predictive mean and variance functions of GPR. Another advantage of GPR is the availability of modern kernel functions; therefore, hyperparameters can be adapted efficiently by boosting the marginal likelihood in the training set (Verrelst et al., 2016). Gaussian processes have been used in a variety of applications such as model approximation, experiment design, and multivariate regression in different scientific disciplines; however, to our knowledge, no previous applications of GPR to DMS prediction have been documented.

The improved long-term (1998–2021) high-resolution seawater DMS concentration data is expected to positively impact the estimated DMS flux into the NA atmosphere and may potentially aid in the parametrization of biogenic sulfur aerosol concentrations in the atmosphere. The study is structured as follows. We begin by using a simple linear regression approach to investigate the relationships between observed DMS concentrations from NA cruises and various environmental parameters potentially acting as DMS predictors. We then perform a multilinear regression to assess each variable's contribution to DMS variance and select the significant predictors to train the machine learning models. As a result, we train and test the GPR model with DMS measurements and the selected most relevant environmental parameters. Furthermore, a comparison with the previous empirical and machine learning algorithms is done to assess if applying GPR improves the NA seawater DMS predictions or not. Then we extend the developed model to obtain gridded fields of daily (0.25° × 0.25°) DMS distributions. The spatial variability and the annual evolution of the DMS in the NA from 1998 to 2021 are investigated using cluster analysis. Lastly, we calculate sea-to-air F_{DMS} over the NA domain and quantify their spatio-temporal variations.

2. Materials and methods

2.1. Data sources

In this study, a combination of in-situ observations, satellite measurements, and modelled data was used to generate daily DMS and F_{DMS} time series over the NA domain from 1998 to 2021. All data sources and their spatial-temporal resolutions are listed in Table 1. The in-situ surface ocean DMS concentrations were obtained from the Global Surface Seawater DMS Database (Pacific Marine Environmental Laboratory, PMEL (Kettle et al., 1999); last access: 14 March 2022) and the North Atlantic Aerosols and Marine Ecosystems Study (NAAMES) (Behrenfeld et al., 2019). We restricted DMS measurements between 1998 and 2021 in the NA domain from 35° N to 66° N and from 55° W to the prime meridian. The cruises tracks are shown in Fig. S1, and the temporal coverages are listed in Table S1. A total of 9484 data points were obtained. The DMS data were binned into a daily resolution and averaged into 0.25° × 0.25° (~28 km) grid cells, eventually, resulting in a total of 2236 data points. The binned dataset was used in the training and validation of the machine learning models (Section 2.3).

The chlorophyll-*a* concentration (CHL) and diffuse attenuation coefficient (K_d) were derived from the Copernicus-GlobColour Satellite Observations at level L4 – daily (0.042° × 0.042°) resolutions. Daily satellite-based time-series of sea surface temperature (SST) were obtained from the ESA Climate Change Initiative project at 0.05° × 0.05° spatial resolution (Merchant et al., 2019). The oceanic mixed layer depth (MLD) and sea surface salinity (SSS) were extracted from the EU Copernicus Marine Environment Monitoring Service (CMEMS) global ocean physics reanalysis at a daily 0.083° × 0.083° resolution. The above-mentioned variables were binned and post-processed to 0.25° × 0.25° resolution.

Nutrient data including sea surface nitrate (NO_3), phosphate (PO_4) and silicate (Si) were used from the CMEMS global ocean biogeochemistry hindcast. The biogeochemical hindcast provides daily 3D fields for the period 1993-ongoing at 0.25° × 0.25° horizontal resolution and on 75 vertical levels. Nutrients were considered in the linear/ multilinear regression and machine learning models because they can affect phytoplankton distributions and thus DMSP production and its subsequent cleavage to DMS (Wang et al., 2015; Zindler et al., 2014).

To cover the 24 years of the present study, the daily photosynthetically available radiation (PAR) available from NASA Ocean Color was downloaded from three products: SeaWiFS (1998–2002), MODIS-Terra (2001 − 2021), and MODIS-Aqua (2003 − 2021). SeaWiFS has 9 km while MODIS has 4 km spatial resolution, both are L3 products. The data were combined and binned to 0.25° × 0.25°, and then linear interpolation was used to fill in the missing values (represent about 5.6 ± 5.9 % of the whole studied domain) and re-processed them as L4 data.

2.2. Simple and multilinear regression

The methodological framework applied in the present study is summarized in Fig. 1 including simple linear regression (A), multilinear regressions (B), and the generation/ cross-validation and testing of machine learning models (C). We conducted the simple regression fit on the daily binned 0.25° × 0.25° DMS datasets from PMEL and NAAMES cruises and possible predictors to explore the predictive skill of each variable separately. Predictors such as CHL, SST, MLD, PAR, SSS, NO_3, PO_4 and Si were assessed which consist of a total of 2236 simultaneous pairs of data points. To reduce the dynamic range of these parameters, we log-transform the DMS and predictors. To avoid losing data with SST less than or equal to 0 °C, the absolute SST was used. The corresponding predictors were standardized to their *z* score, where each predictor is centered to have mean = 0 and scaled to have standard deviation = 1,

Table 1
The main data, sources, and their spatial-temporal resolutions used in the present study. Data processing levels are L4 except PAR (L3). All variables were binned and post-processed to daily averaged– 0.25° resolution. Temporal coverage of the data from 1998 to 2021 in the north Atlantic domain (35–66°N & 55°W – 0°E).

Variable	Spatial resolution	Temporal resolution	Source	Unit
DMS	In situ measurements from ocean cruises		PMEL http://saga.pmel.noaa.gov/dms/ NAAMES https://doi.org/10.5067/SeaBASS/NAAMES/DATA001	μmol m^{-3}
Chlorophyll-*a* (CHL)	0.042° × 0.042°	Daily	Copernicus-GlobColour Satellite Observations https://doi.org/10.48670/moi-00281	mg m^{-3}
Diffuse attenuation coefficient (K_d)	0.042° × 0.042°	Daily	Copernicus-GlobColour Satellite Observations https://doi.org/10.48670/moi-00281	m^{-1}
Sea surface temperature (SST)	0.05° × 0.05°	Daily	ESA Sea Surface Temperature Climate Change Initiative v2.1 https://cds.climate.copernicus.eu	°C
Mixed layer depth (MLD) Sea surface salinity (SSS)	0.083° × 0.083°	Daily	CMEMS Global Ocean Physics Reanalysis https://doi.org/10.48670/moi-00021	m g kg^{-1}
Sea surface nitrate (NO_3) Sea surface phosphate (PO_4) Sea surface silicate (Si)	0.25° × 0.25°	Daily	CMEMS Global Ocean biogeochemistry hindcast https://doi.org/10.48670/moi-00019	mmol m^{-3}
Photosynthetically available radiation (PAR)	9 km (SeaWiFS) 4 km (MODIS)	Daily	NASA Ocean Color SeaWiFS (1998–2002), MODIS-Terra (2001–2021), and MODIS-Aqua (2003–2021) https://oceancolor.gsfc.nasa.gov	mol photons m^{-2} d^{-1}
Neutral wind speed at 10 m (W_{10n})	0.25° × 0.25°	Hourly	ECMWF ERA5 https://doi.org/10.24381/cds.adbb2d47	m s^{-1}

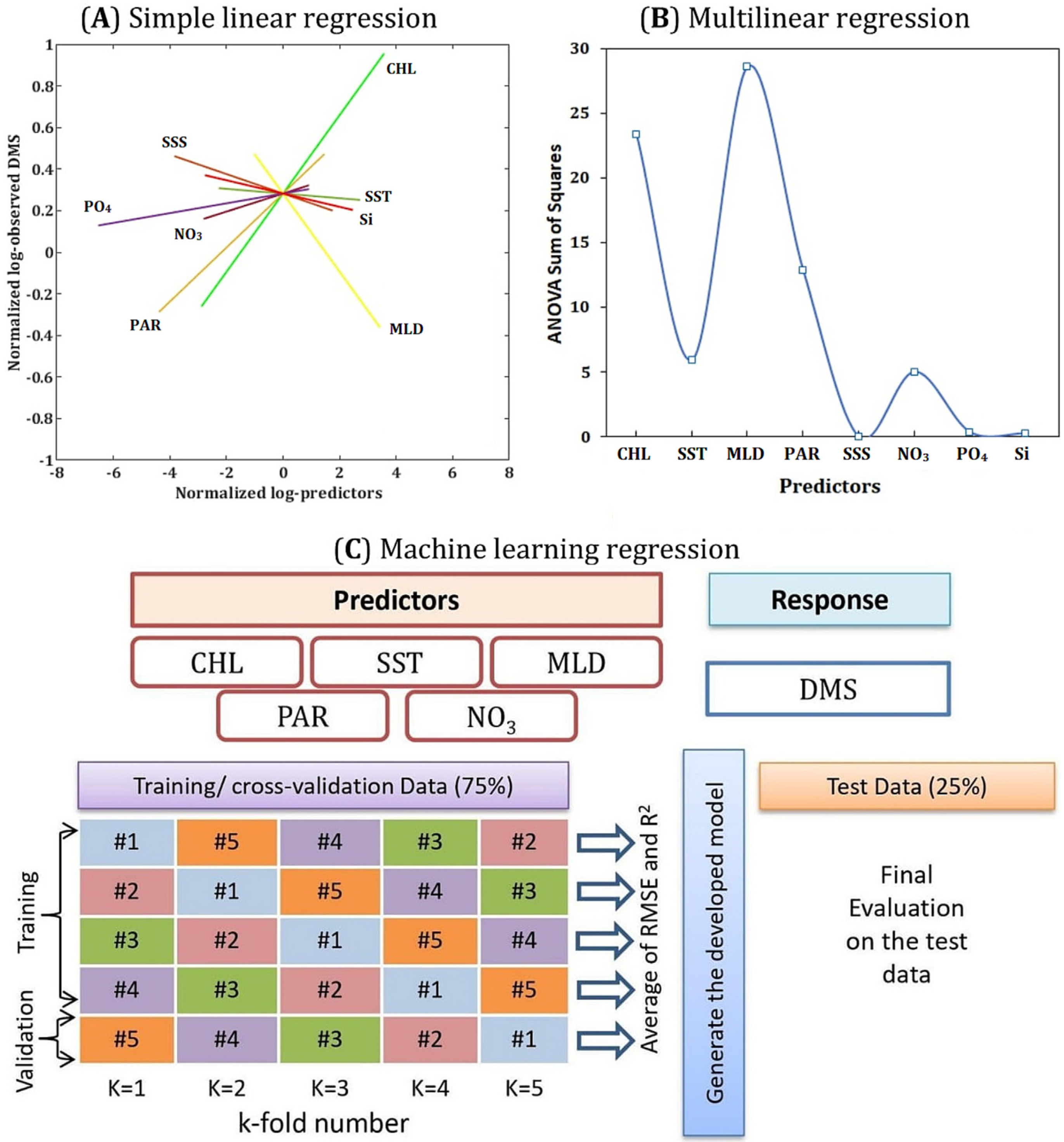


Fig. 1. Flowchart of the methodological frame applied in the present study. (A) and (B) represents the results from simple and multilinear regressions. (C) Diagram of the GPR model generation and validation details, including cross-validation and testing.

as recommended by Wang et al. (2020). The linear regression function was applied to each pair to fit a first-degree polynomial.

From simple linear to multilinear regression, we applied, in the first step, the regression using a total of the potential eight DMS predictors: CHL, SST, MLD, PAR, SSS, NO_3, PO_4 and Si. In the second step, we applied the multilinear regression model by eliminating one predictor each time to assess each predictor's contribution to the explained DMS variance. The contribution to R^2 of each independent variable is the reduction in the total R^2 when that variable is omitted. Such multilinear regressions between predictors and DMS were performed on logarithmically transformed variables, allowing for the non-linear relationships to be taken into account.

2.3. Gaussian process regression (GPR)

Based on the results of simple and multilinear regressions, we used five independent predictors (CHL, SST, MLD, PAR, and NO_3), which conceivably contribute to DMS variance, to build up the GPR model for DMS prediction in the NA. The workflow diagram of the model generation and evaluation is shown in Fig. 1C. First, the dataset, containing DMS concentration data and the corresponding predictors data for each point of the available grid of observations, was split randomly into two subsets: a set for model training and cross-validation (75 % of the total points; $n = 1676$) and a set for model testing and evaluation (25 %; $n = 560$). A 5-fold cross-validation strategy was used which means that the training dataset was further divided into 5 groups, or folds, of approximately equal size. At each trial, a unique group is taken as a holdout or validation dataset and the remaining four groups as training data, then the model is fit on the training set and evaluated on the validation set. The average evaluation measures (accuracy) of the five iterations were reported. Ultimately, the developed model was further evaluated on the test data, which was not used in the model generation, to better evaluate the repeatability of the model on a new dataset. When developing the GPR, different base kernel (covariance) functions, namely exponential, Matern 5/2, squared exponential, and rational quadratic (Asante-Okyere et al., 2018) were assessed to determine the optimal covariance function that could produce reliable predictions closely related to the observed DMS. For more information on GPR, the reader is referred to the MATLAB help documentation (https://www.mathworks.com/help/stats/fitrgp.html).

2.4. The existing DMS retrieval algorithms

We applied previously published empirical algorithms for calculating DMS seawater concentration to test the performance of our GPR model. Equations from Simo and Dachs (2002), Vallina and Simo (2007), and Gali et al. (2018) (hereafter referred to as SD02, VS07, and G18, respectively) were used to predict DMS in the NA during the studied period.

SD02 parameterized DMS as a linear function of CHL/MLD when CHL/MLD ≥ 0.02, which explains as much as 84 % of the global DMS variance grouped by roughly 10° latitudinal bands, as well as a logarithmic negative relationship between DMS and MLD when the CHL/MLD < 0.02, which explains 68 % of the variance of the subset DMS. The equations are as follows:

$$\boldsymbol{DMS} = 55.88 \times \left(\frac{\boldsymbol{CHL}}{\boldsymbol{MLD}}\right) + 0.6 \ \text{ for } \ \boldsymbol{CHL}/\boldsymbol{MLD} \geq 0.02 \tag{1}$$

$$\boldsymbol{DMS} = -\boldsymbol{ln}\ (\boldsymbol{MLD}) + 5.7 \ \text{ for } \ \boldsymbol{CHL}/\boldsymbol{MLD} < 0.02 \tag{2}$$

VS07 parameterized DMS concentration according to the strong linear relationship with solar radiation dose (SRD) in the coastal northwestern Mediterranean and generalized the linear coefficients to the global open ocean divided into grids of 10° latitude by 20° longitude as:

$$\boldsymbol{DMS} = 0.492 + 0.019 \times \boldsymbol{SRD} \tag{3}$$

In the 14 grid boxes which cover the global ocean, the linear regression of Eq. (3) yielded an R^2 value of 0.95. In the present study, the SRD in Eq. (3) was calculated using the following formula:

$$\boldsymbol{SRD} = \frac{\boldsymbol{I}}{\boldsymbol{K_d} \times \boldsymbol{MLD}} \times \left(1 - \boldsymbol{e}^{-\boldsymbol{K_d} \times \boldsymbol{MLD}}\right) \tag{4}$$

where I is the average intensity of surface radiation (W m^{-2}) and was calculated from PAR using a conversion factor by Morel and Smith (1974). K_d is the average diffuse attenuation coefficient of downwelling radiative flux in seawater (Table 1).

The most recent algorithm for DMS retrieval was introduced by Gali et al. (2018). It proceeds through two steps, calculation of DMSP in seawater and then retrieving DMS. The DMSP algorithm (Gali et al., 2015) switches between two different equations depending on the quotient between euphotic zone depth (Z_{eu}; 1 % light penetration) and MLD. This quotient is used to distinguish between stratified water column if Z_{eu}/MLD ≥ 1 (Eq. (5) is used) and mixed water column if Z_{eu}/MLD < 1 (Eq. (6) is applied).

$$\begin{aligned}\boldsymbol{log}_{\ 10}(\boldsymbol{DMSP}) = {} & 1.70 + 1.14\lambda + 0.44\,\lambda^2 \\ & + 0.063\,\boldsymbol{SST} - 0.0024\,\boldsymbol{SST}^2\end{aligned} \tag{5}$$

$$\boldsymbol{log}_{\ 10}(\boldsymbol{DMSP}) = 1.74 + 0.81\lambda + 0.60\ \boldsymbol{log}\ \ (\boldsymbol{Zeu}/\boldsymbol{MLD}) \tag{6}$$

where $\lambda = \log_{10}$ (CHL). CHL in mg m^{-3}, SST in °C and Z_{eu} and MLD in m. The Z_{eu} is estimated as a function of CHL following Morel et al. (2007) as:

$$\boldsymbol{log}_{\ 10}(\boldsymbol{Zeu}) = 1.524 - 0.436\lambda - 0.0145\,\lambda^2 + 0.0186\,\lambda^3 \tag{7}$$

The sea-surface DMS concentrations (μmol m^{-3}) are estimated from DMSP and PAR by applying Eq. (8). This algorithm accounted for 56 % of the monthly DMS variance binned globally into 5° × 5°.

$$\boldsymbol{log}_{\ 10}(\boldsymbol{DMS}) = -1.237 + 0.578\ \ \boldsymbol{log}_{\ 10}(\boldsymbol{DMSP}) + 0.018\ \boldsymbol{PAR} \tag{8}$$

For direct comparison to the GPR ensemble performance, the three aforementioned algorithms (SD02, VS07, and G18) were applied in the NA domain to generate daily DMS concentrations at the 0.25° × 0.25° spatial resolution (For more information, see Section 3.2).

2.5. K-means + + clustering

Cluster analysis is an unsupervised machine learning technique that divides a set of data points into groups or clusters to maximize the variance between clusters while minimizing variance within each cluster. In this work, the k-means + + clustering algorithm (https://www.mathworks.com/help/stats/kmeans.html) was applied to identify regions with similar patterns in the seasonal cycle of surface DMS concentrations over the NA ocean. The daily predicted DMS data from GPR were used over the 1998–2021 period at each pixel (spatial resolution of 0.25° × 0.25°) of the NA basin. To facilitate comparison between pixels, each time vector for each pixel was normalized by the maximal and minimal DMS value to be scaled between zero and unity. Due to the low-incident sun angle in winter, the CHL time series could not be measured from satellites at high latitudes which hinders DMS calculations. For this reason, only periods from March to September were considered in this analysis. K-means + + is a smart centroid initialization method (Arthur et al., 2007) in which the first centroid is picked randomly from the data points and the remaining centroids are chosen based on the maximum squared distance. Such an initial random centroid selection in k-means + + achieves faster convergence to a lower sum of the within-cluster sum of squares point-to-cluster-centroid distances than in the classic k-means Lloyd's algorithm (Lloyd, 1982), and improves the quality of the final solution (Arthur et al., 2007).

As an initial step of clustering, the optimal number of clusters, *k*, within DMS dataset was defined by applying the advanced Elbow point discriminant method (Shi et al., 2021). The method proposes a statistical metric, cosine of interaction angle, from which an optimal number of clusters can be estimated. Given that a set of time series $X = x_1, x_2, \ldots x_n$, where each series represent a time-vector of a certain parameter at a pixel in the NA domain. Clustering aims to partition the *n* time series into the number of clusters $k = 1, 2, \ldots L$ ($L \leq n$). The centroids corresponding to the cluster *k* are defined as $\mu_1, \mu_2, \ldots \mu_k$. For example, if $k = 5$, this means that *X* is divided into 5 groups which have centroids μ_1, μ_2, μ_3, μ_4 and μ_5. The within-cluster sum of the squared error (*SSE*) is the sum of the square Euclidean distance between each data point belonging to the same cluster and its cluster centroid. The total *SSE* is given by:

$$\boldsymbol{SSE_{total}} = \sum_{k=1}^{L} \sum_{i=1}^{n} (\boldsymbol{x_i} - \boldsymbol{\mu_k})^2 \tag{9}$$

The mean distortion (*MD*) of the dataset *X* of *n* pixels is given by:

$$\boldsymbol{MD} = \boldsymbol{SSE_{total}}/\boldsymbol{n} \tag{10}$$

The next step in the method is rescaling the vector *MD* to span from 0 to 10 using Eq. (11).

$$\boldsymbol{SMD} = \frac{\boldsymbol{MD} - \ \min\ (\boldsymbol{MD})}{\max\ (\boldsymbol{MD}) - \ \min\ (\boldsymbol{MD})} \times \mathbf{10} \tag{11}$$

For three successive clusters a, b and c, where a, b, c □ 1, 2, …, L, Eq. (12) can be used to calculate the Euclidean distance between each two of them.

$$\boldsymbol{E_{ab}} = \sqrt{(\boldsymbol{SMD_a} - \boldsymbol{SMD_b})^2 + (\boldsymbol{a} - \boldsymbol{b})^2} \tag{12}$$

The Elbow interaction angle α_b is given by:

$$\boldsymbol{\alpha_b} = \ \mathbf{cos}^{\ -1}\left(\frac{\boldsymbol{E_{ab}^2} + \boldsymbol{E_{bc}^2} - \boldsymbol{E_{ac}^2}}{\mathbf{2} \times \boldsymbol{E_{ab}^2} \times \boldsymbol{E_{bc}^2}}\right) \tag{13}$$

The smallest α indicates the optimum Elbow point in the space of $\alpha = \alpha_1, \alpha_2, \ldots \alpha_{L-2}$ and the corresponding k is the estimated potential optimal cluster number for the analyzed dataset.

2.6. Sea-to-air DMS flux

Daily sea-to-air DMS fluxes (F_{DMS}; μmol m^{-2} d^{-1}) from 1998 to 2021 were calculated using surface ocean DMS concentrations and the ECMWF-ERA5 (Hersbach et al., 2020) reanalysis of 10 m neutral wind speed by applying Eq. (14).

$$\boldsymbol{F_{DMS}} = \boldsymbol{k_{DMS}} \times \boldsymbol{DMS} \tag{14}$$

DMS in Eq. (14) represents the values in sea surface water concentration (μg m^{-3}). The assumption is based on DMS in the surface ocean being strongly supersaturated to that in the overlying atmosphere (Wang et al., 2020), and so the air-sea concentration difference ΔDMS is almost equal to seawater DMS. The gas transfer velocity (k_{DMS}) is estimated as a function of neutral wind speed following Goddijn-Murphy et al. (2012) parametrization.

$$\boldsymbol{k_{DMS}} = (\mathbf{2.1}\boldsymbol{W_{10n}} - \mathbf{2.8})\left(\frac{\boldsymbol{Sc_{DMS}}}{\mathbf{660}}\right)^{-\mathbf{0.5}} \tag{15}$$

where W_{10n} is the neutral wind speed (m s^{-1}) at 10 m above the sea surface. Sc_{DMS} is the Schmidt number (diffusivity of DMS through seawater), which is dependent upon sea surface temperature (SST) and is calculated by Saltzman et al. (1993) as follows:

$$\boldsymbol{Sc_{DMS}} = \mathbf{2674.0} - \mathbf{147.12}\ \boldsymbol{SST} + \mathbf{3.726}\ \boldsymbol{SST}^{\mathbf{2}} - \mathbf{0.038}\ \boldsymbol{SST}^{\mathbf{3}} \tag{16}$$

Eq. (15) is valid only when $W_{10n} > 1.33$, otherwise, it gives negative values. Negative $\boldsymbol{k_{DMS}}$ are replaced by applying linear interpolation.

3. Results and discussion

3.1. Linear regression and predictors selection

The linear regression fit lines between daily binned observed DMS and possible predictors (CHL, SST, MLD, PAR, SSS, NO_3, PO_4 and Si) individually are shown in Fig. 1A (frequency distributions are shown in Fig. S2), whereas slopes and R^2 values are summarized in Table 2. The strongest predictors of DMS in seawater are CHL and MLD ($R^2 = 0.21$, $n = 2236$). The positive correlation between CHL and DMS (slope = 0.19) can be attributed to the fact that DMSP, the precursor of DMS, derives from biological processes, which are tracked by the abundance of CHL in surface seawater. The deepening of the ocean MLD tends instantaneously to reduce the DMS at the sea surface as a result of the vertical mixing and of the reduced exposure to radiation, resulting in a reverse MLD-DMS relationship (slope = −0.19). Gali et al. (2018) and Wang et al. (2020) reported a quasi-similar R^2 value of CHL-DMS using in situ global data, while the R^2 is lower using binning monthly 5° × 5° data ($R^2 = 0.14$) (Gali et al., 2018). The third strongest predictor for DMS is PAR which can describe 10 % of DMS variance in a positive relationship. Notably, there is a strong similarity between the linear coefficients and R^2 of the globally in-situ data (Wang et al., 2020) and our daily binned satellite data obtained in the NA.

Table 2
Simple linear regression between observed DMS and possible predictors. The R^2 and slopes values are for log-log space and normalized data as described in the text.

Parameter	CHL	SST	MLD	PAR	SSS	NO_3	PO_4	Si
R^2	0.21	0.00	0.21	0.10	0.01	0.01	0.00	0.01
Slope	0.19	−0.01	−0.19	0.13	−0.05	0.04	0.02	−0.03

The applied multilinear regression model (Table 3), which employs a combination of the eight predictors, outperforms the linear regression model in terms of predictive ability (total R^2 value of 0.39, which is higher than that of any of the linear models). In addition, Table 3 shows that three predictors (MLD, CHL and PAR) explain up to 32 % of the variability in DMS. The SST and NO_3 can explain the rest of the total variance while the SSS, PO_4 and Si have a negligible impact. To support this, we perform the analysis of variance (ANOVA) on the implemented multilinear regression model of the eight predictors. The ANOVA sums of squares (Fig. 1B), which show how much of the variance is explained by each component of the decomposition, reveal similar contributions from the tested predictors as the multilinear regression. No statistically significant ($p < 0.05$) contribution from SSS, PO_4 and Si is observed (Table S2). For these reasons, we applied the machine learning model (GPR) using the 5 predictors (MLD, CHL, PAR, SST and NO_3) that contribute significantly to the total variance.

3.2. GPR training/ cross validation and testing/ evaluation

A 5-fold cross-validation strategy was used to develop the GPR model (Fig. 1C). The 75 % of all the data used in model training and cross-validation is subdivided into 5 subgroups. The iterations are repeated five times and in each iteration, data from 4-folds is used in model training while the data in the remaining fold is used for validation, such that a different fold is selected as a validation set each time (Fig. 1C). As a primary step, we assessed four base kernel (covariance) functions which are exponential, Matern 5/2, squared exponential, and rational quadratic, utilizing the same 5-fold plan. Then, the four models executed by different kernel functions were extended to the test data. The motivation was to identify the best-performing (optimal) kernel function suitable for accurately predicting DMS concentrations. The average evaluation measures of the five trials corresponding to the 5-fold strategy are summarized in Table S3. While the four applied kernel functions have quasi-similar measures, the best optimal GPR model for predicting DMS is the exponential covariance function. The GPR model based on the exponential covariance function achieves the highest R^2 value of 0.67 and the least root mean square error (RMSE) of 0.24 for the training data. When extending to the test data, R^2 and RMSE reach 0.71 and 0.21, respectively. Therefore, the covariance exponential function of GPR was selected as the optimal regressor for further analysis throughout this study. The model was saved and run to obtain gridded fields of daily DMS (1998–2021) distributions over the NA.

Fig. 2A displays the comparison between observed and predicted DMS by the developed GPR model. When compared to simple linear (Table 2)

Table 3
Multilinear regression of DMS as a function of predictors. The contribution to R^2 for each of the independent variables is the decrease in total R^2 with that variable omitted. The sums of individual contributions of R^2 are adjusted to equal the total R^2.

	Total R^2	RMSE	Normalized contribution to R^2							
			CHL	SST	MLD	PAR	SSS	NO_3	PO_4	Si
DMS	0.39	0.32	0.12	0.03	0.14	0.06	0.00	0.03	0.00	0.00

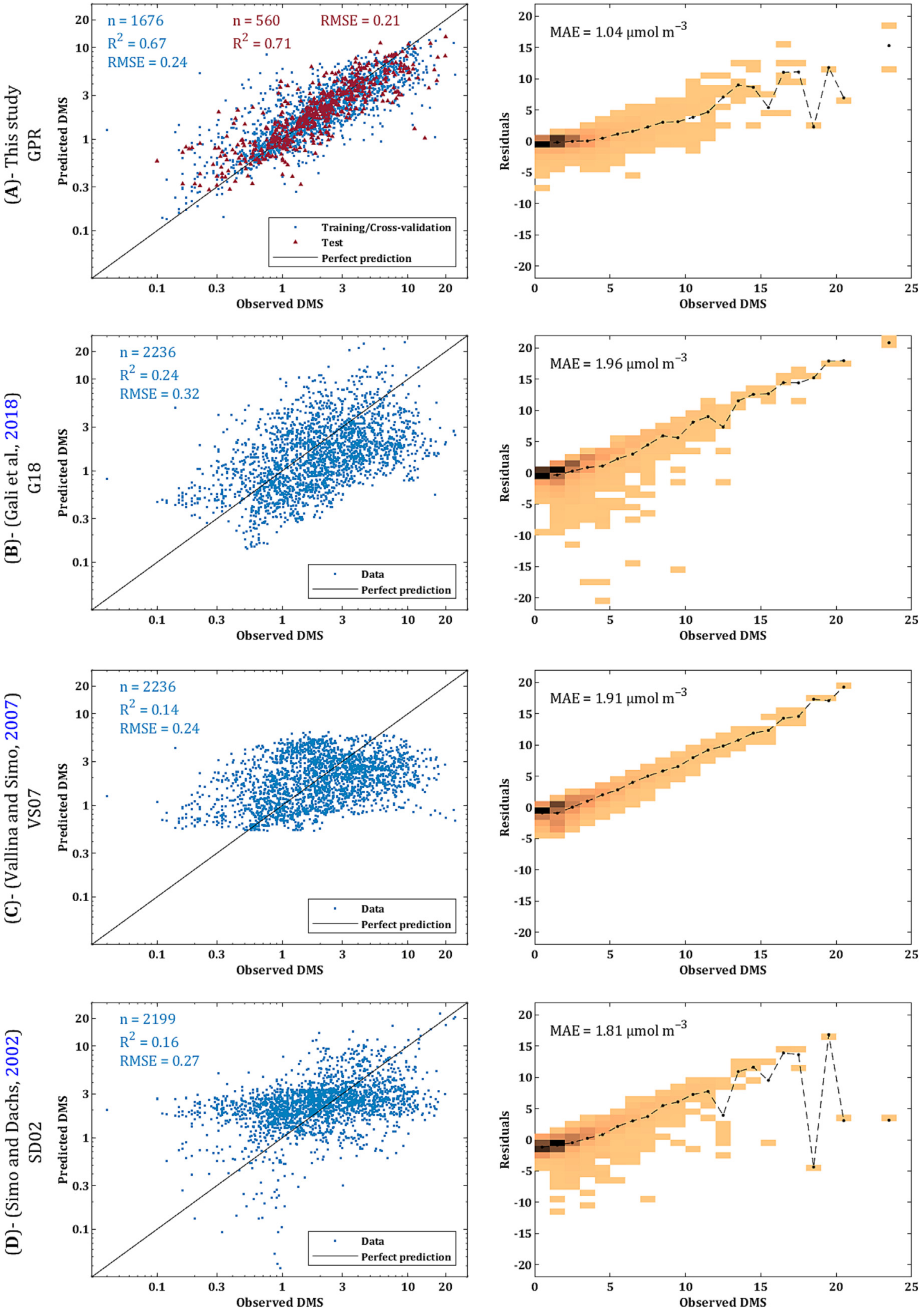


Fig. 2. Predicted versus observed DMS (μmol m^{-3}): (A) gaussian process regression presented in this work, (B—D) previous estimates based on linear and multilinear regressions (see Section 2.4). GPR performance (R^2 and RMSE) are computed for the cross-validation and test datasets. The right panel represents joint probability histograms where darker colors indicate higher probability; the colors are normalized so that the sum of total pixels is 1 in each plot. The dashed black lines represent the change of DMS residual errors (observed–predicted) in each bin. MAE is the mean absolute error.

Table 4
Summary of performance measures of ANN and RFR in predicting DMS in the NA.

		ANN type					RFR
		Narrow	Medium	Wide	*Bi*-layered	Tri-layered	
	Number of fully conneted layers	1			2	3	
	Layer size	1st = 10	1st = 25	1st = 100	1st = 10 2nd = 10	1st = 10 2nd = 10 3rd = 10	
Training/cross-validation	RMSE	0.303	0.288	0.295	0.288	0.288	0.260
	R^2	0.47	0.52	0.50	0.52	0.52	0.61
Test	RMSE	0.255	0.259	0.254	0.252	0.260	0.233
	R^2	0.58	0.57	0.59	0.59	0.57	0.65

and multilinear (Table 3) regression models, it is clear that GPR can reconstruct the observations with a markedly higher R^2 value (0.71 against 0.39 of the multilinear model), which means that the selected machine learning approach captures much more of the observed DMS variability. To benchmark the performance of our GPR model, we compared the predictive skills of three existing empirical DMS algorithms (G18, VS07, and SD02) applied to the daily binned NA data against the GPR. We found that the G18, VS07, and SD02 did not accurately predict the NA DMS at high temporal and spatial resolutions (Fig. 2B-D). The best performing of the current existing numerical algorithms was G18 where $R^2 = 0.24$ was achieved, but quantitatively RMSE was higher than VS07 and SD02. The high RMSE may be due to the use of the global coefficient (Eqs. (5) and (6)) in DMSP retrieval (Gali et al., 2015) which may not be suitable on a regional scale. It is worth highlighting that the applied multilinear regression in this study performs better than G18, even using the same predictors (R^2 moved from 0.24 [G18] to 0.35 [Table 3]).

The joint probability histograms between observed DMS and the residuals (observed – predicted) are used to verify the variance of residual errors around zero (right panel of Fig. 2). The GPR histogram shows that the residual errors are mostly centered around zero (dashed black line in the right panel of Fig. 2), while G18, VS07 and SD02 skewed toward positive residuals mainly at high DMS values. Quantitatively, the DMS tested/ cross-validated data points predicted by GPR have mean absolute error (MAE) equal to 1.04 μmol m^{-3}. Higher variability in residual errors is observed for the other existing algorithms. For instance, G18, VS07, and SD02 yield mean absolute error (MAE) equal to 1.96, 1.91, and 1.81 μmol m^{-3}, respectively (Fig. 2, right panel). Accordingly, compared to the other previously published empirical approaches, the GPR model significantly improves the representation of NA DMS variability, achieving significantly higher prediction accuracy and lesser errors. Aware that the GPR model could be biased due to the uneven distribution of in situ DMS measurements, we constructed the joint probability histograms of DMS residuals versus latitude and longitude, separately. The histograms (Fig. S3) show that DMS residuals are mostly around zero without any tendency, not longitudinally nor latitudinally.

Eventually, we evaluated the use of the GPR model to predict DMS by excluding, from the analysis, the NO_3 data, which are calculated using numerical modeling techniques. In this way, we considered only predictors that can be obtained directly (CHL, SST, and PAR) or indirectly (MLD) from satellite observations. Excluding NO_3 has no substantial effect on GPR output. On the test data, the model can explain 0.68 of the observed DMS variance (RMSE = 0.22) instead of $R^2 = 0.71$ and RMSE = 0.21.

3.3. Comparison of GPR with ANN and RFR

The GPR model is then compared to previously applied machine learning algorithms: ANN (Bell et al., 2021; McNabb and Tortell, 2022; Wang et al., 2020) and RFR (McNabb and Tortell, 2022), which were recently used for seawater DMS prediction. To make a proper comparison, the same predictors for DMS used in GPR were also used to train the ANN and RFR algorithms and the same subset for testing was used to ensure the repeatability of the models. We trained various types of ANN as single-layer, bi-layered, and tri-layered neural networks. Based on the extension to the test dataset, the best performing ANN model is the bi-layered model with an R^2 value of 0.59 and an RMSE of 0.25, as shown in Table 4. The RFR method provides slightly better results ($R^2 = 0.65$; RMSE = 0.23), compared to the individual ANN models. In comparison, the proposed GPR model outperforms ANN and RFR algorithms over the NA domain. The comparison shows that a fraction of 29 %, 35 % and 41–43 % of the DMS variance could not be captured by GPR, RFR and different ANN methods, respectively. The comparative analysis suggests that the proposed GPR is a promising approach to obtaining reliable DMS concentration values suggesting that it may also be equally successful in other oceanic regions or, perhaps, even on a global scale.

3.4. Monthly DMS distributions

The best GPR model based on the exponential function was used to create gridded fields of daily DMS distributions across the NA covering the entire period of 1998–2021. The DMS climatology from GPR was compared geographically to the DMS from G18 (the best in previous arithmetic methods) and the in-situ measurement-based climatology from Lana et al. (2011); hereafter referred to as L11. The maps are presented in Fig. 3. The statistical comparison of the three climatology products over the NA domain is summarized in Table 5.

The maps in Fig. 3A display the climatological monthly mean sea surface DMS concentrations predicted by the GPR model from March to September (1998–2021). We restrict this discussion to the spring-summer months due to the impossibility of calculating winter DMS concentrations at high latitudes. Winter DMS distribution maps are presented in Fig. S4. Relatively high DMS concentrations characterize the northern part of the studied domain, essentially in summer (Jun-Aug). The spatial features of the DMS concentration are mostly related to the distribution of CHL and the CHL/MLD ratio (Fig. S5). This is reasonable since we show that both CHL and MLD are the best predictors for observed DMS. The seasonal trend of DMS concentration is evident: the increase starting in March and peaking at about 2.94 ± 1.14 μmol m^{-3} in July (Table 5) followed by a gradual decrease in September. The spring-summer (March–September) mean DMS concentration over the domain is 2.24 ± 0.48 μmol m^{-3}.

The maps in Fig. 3B show the G18 DMS climatology. Note that the resolution of the data used to construct the G18 climatology is daily (0.25° × 0.25°) during the same investigated period 1998–2021. Our comparison (GPR vs. G18) indicates that G18 show lower concentrations than GPR during March, August, and September, mainly over the northern areas of the studied domain. From April to July, G18 shows mostly higher concentrations over the whole NA. Most notably, G18 does not capture the zonal decrease of DMS concentration toward the equator during summer (Jun-August). Such low DMS concentrations below 50°N are evident by the GPR and consistent with the latitudinal change of in situ DMS measurements (Fig. S6) and are further supported by the spatial distributions of CHL and CHL/MLD (Fig. S5). A possible explanation may be that Eq. (8), used by G18, results in an overestimation of DMS concentration at high PAR (Fig. S5) during the summer season.

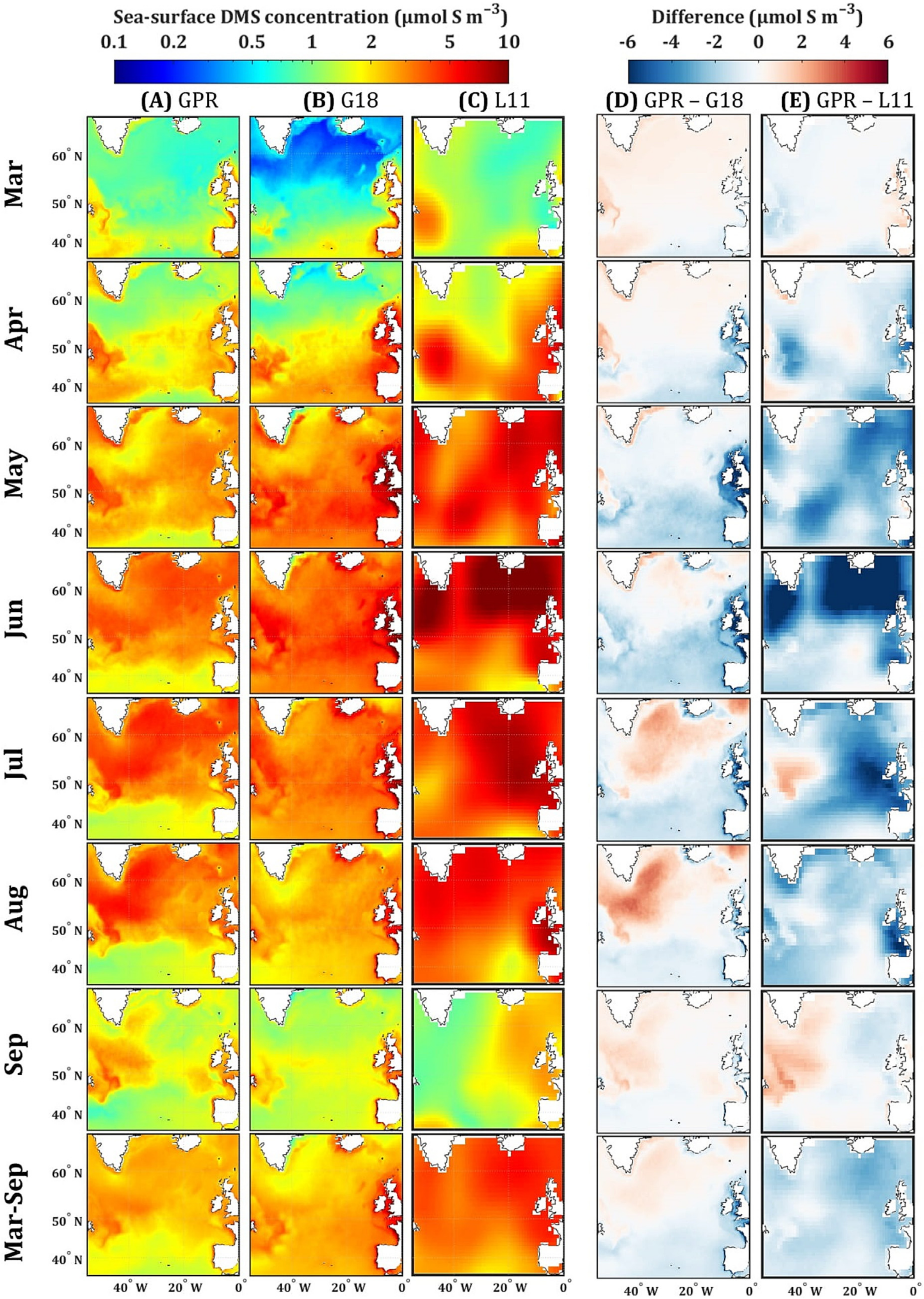


Fig. 3. Spatial comparison of climatological monthly mean sea surface DMS concentration based on (A) GPR over 1998–2021, (B) G18 over 1998–2021, and (C) The L11 climatology from the extrapolation of gridded in situ measurements. (D) The difference between the GPR and G18 climatology. (E) The difference between the GPR and L11. The pixels resolution of (A), (B), and (D) is 0.25° × 0.25° while (C), and (E) have 1° × 1° spatial resolution.

Table 5
Statistics of sea surface DMS concentration comparisons among the GPR, G18, and in situ measurements (L11) climatology products over the NA waters. Percentages in brackets indicate the relative difference (%) with respect to G18 and L11 climatology.

	GPR	G18	L11	GPR – G18	GPR – L11
	Mean ± spatial standard deviation (μmol m^{-3})			Spatial mean difference (μmol m^{-3})	
Mar	1.17 ± 0.49	0.92 ± 0.70	1.34 ± 0.57	0.25 (+26 %)	−0.18 (−13 %)
Apr	1.84 ± 0.65	2.15 ± 1.23	2.81 ± 1.23	−0.30 (−14 %)	−0.98 (−34 %)
May	2.52 ± 0.61	3.72 ± 1.37	4.70 ± 1.33	−1.19 (−32 %)	−2.18 (−46 %)
Jun	2.86 ± 0.82	4.06 ± 1.26	6.01 ± 3.18	−1.19 (−30 %)	−3.14 (−52 %)
Jul	2.94 ± 1.14	3.41 ± 1.05	5.05 ± 1.92	−0.47 (−14 %)	−2.11 (−42 %)
Aug	2.65 ± 1.16	2.54 ± 0.85	4.28 ± 1.47	0.12 (+04 %)	−1.62 (−38 %)
Sep	1.70 ± 0.56	1.58 ± 0.59	1.60 ± 0.58	0.13 (+07 %)	0.09 (+06 %)
Mar-Sep	2.24 ± 0.48	2.63 ± 0.87	3.68 ± 0.85	−0.38 (−15 %)	−1.45 (−39 %)

The in-situ measurement-based L11 DMS climatology is presented in Fig. 3C. The L11 data are interpolated/extrapolated 1° × 1° monthly mean fields of DMS global climatology based on available oceanic DMS concentration measurements worldwide taken between 1972 and 2009. We cut the climatology in the studied NA domain, while the temporal coverage is necessarily different between GPR and L11. In Jun, the L11 climatology shows a relatively high concentration (6 μmol m^{-3}) on average over the NA domain, whereas the values reach up to 10 μmol m^{-3} in the south of Iceland, the Norwegian Sea, and the Labrador Sea. Unlike GPR and G18 model climatology, DMS concentrations by L11 are unable to describe the key characteristics of contributory predictors such as CHL and CHL/MLD. This is probably due to the impact of high in-situ DMS concentrations having a far-reaching impact on L11 extrapolation.

The maps in Fig. 3D show the difference between the GPR and G18 climatology fields over the NA. The GPR model climatology is about 0.38 μmol m^{-3} lower than the G18 one (the relative difference is −15 %) in the whole NA from March to September (Table 5). The maximum difference occurred in May and June which is 1.19 μmol m^{-3} (accounting for a − 30 % in relative terms), mainly located in the southern sector of the NA domain. In contrast, the GPR model climatology is about 0.25 (+26 %), 0.12 (+4 %), and 0.13 (+7 %) μmol m^{-3} higher than the G18 one in March, August, and September, respectively. The major differences between GPR and G18 distributions are observed in correspondence to very low CHL values, when DMS concentrations from G18 are mostly impacted by the PAR value, resulting in a lower agreement between the CHL and DMS distributions in G18 (Fig. S5). This can be seen particularly in the southern sector of the study area, where GPR predicts a strong latitudinal gradient of the DMS concentration across the 50th parallel which is not represented by G18. Further, high DMS concentration is found by G18 in the coastal continental shelves (e.g., Infront of Ireland, the Irish Sea, and the Bay of Biscay) as a result of extremes of phytoplankton biomass in such regions, which is not observed in GPR.

The maps in Fig. 3E show the difference between the GPR and in situ measurement-based climatology fields (L11). The GPR generally simulates lower DMS concentrations than L11 (relative change is −39 % in the whole NA from March to September; Table 5), except for regions to the south and southeast of Greenland in July and September. The GPR-based climatology is quantitatively closer to the G18-based climatology than the in-situ L11-based climatology over the NA domain, especially away from May and June, as seen by the mean difference in Table 5. The L11 climatology's mean values are consistently higher than the other two products, particularly in May and June. These findings once again emphasize the potential consequences of the L11 climatology's extrapolation bias. However, the three products are fairly comparable in terms of displaying the DMS seasonal cycle, showing high DMS concentrations from May to August.

3.5. DMS cluster analysis

The above findings demonstrate that DMS concentration over the NA waters exhibits a wide spatial variation due to the variation of its independent predictors. The k-means++ algorithm is applied here to find oceanic regions with comparable DMS seasonal patterns. Then, to better understand the magnitude, initiation, and extension time of DMS peaks in each cluster, we investigate the annual cycle of DMS and how it is related to the potential predictors (CHL, SST, MLD, PAR, and NO_3) used in the model. Between March and September, daily predicted DMS data from GPR are used at each pixel in the NA basin (spatial resolution of 0.25° × 0.25°) in this analysis. The optimal number of clusters is set according to the novel Elbow approach by Shi et al. (2021), which performs better than the Calinski-Harabasz index (Caliński and Harabasz, 1974) and the classic Elbow graph (Syakur et al., 2017) in this analysis, setting the optimal cluster number to 7 (see Section 2.5 and Fig. S7 for details).

The distribution of clusters (Fig. 4A) reveals a marked spatial variability in DMS seasonal pattern over the NA during the investigated period. The daily climatological sea surface DMS concentration and associated averaged predictors in each cluster are presented in Fig. 4B. Notably, the DMS seasonal cycle and the patterns of the predictors vary between the different clusters (Fig. 4). DMS concentrations with large seasonal variability (large standard deviation and high mean values as summarized in Table S4) are located mainly toward the North of the studied domain represented by clusters 1 and 7. In such oceanic regions, DMS and CHL increase simultaneously from May, then DMS concentration continues growing, while CHL remains stable or declining during summer (June–August). Geographically, cluster 6 appears to be representative of the center of the subpolar gyre. It is characterized by a high DMS concentration for 3 months, from May to July. A common feature observed in clusters 1, 6 and 7 is the sharp seasonal variation in MLD from deep in the early spring to shallow in the summer and fall, accompanied by an excess of nutrients in the early Spring. This can be better seen by looking at the data in Table S4 showing that the range of MLD is the largest in those clusters. In those clusters, the time extent of the DMS peak is longer compared to the other ones. This evidences that nutrients from deep water upwelled by strong vertical mixing are necessary to increase the phytoplankton activity and initiate the increase of DMS; later on the enhanced PAR and the shoaling of the MLD stress phytoplankton cells allowing the persistence of high DMS concentrations (Becagli et al., 2013; Simo and Dachs, 2002; Vallina and Simo, 2007). These results are in accordance with previous findings reporting that seawater vertical mixing may affect the duration and magnitude of the DMS concentration peaks also by creating stress conditions in the phytoplankton cells when they are trapped in a shallow surface mixing layer. This favors the production and release of significant amounts of DMS in seawater (Kwint and Kramer, 1995; Laroche et al., 1999; Zhuang et al., 2011) and reduces its dilution. Indeed, a recent study (Diaz et al., 2021) showed that marine organisms in shallow mixed layers, in late spring in the NA ocean, suffer from high levels of oxidative stress. This aspect may contribute to the prolonged DMS concentration peak until late summer.

In the South side of the studied domain, clusters 4 and 5 represent the region between 35° and 45° N, which have the minimum phytoplankton biomass abundance and are characterized by a single spring bloom. This is related to the ocean circulation in the NA (Friedland et al., 2016). In this region, the bloom starts early in spring and propagates northward driven by the NA drift current (Siegel et al., 2002), resulting in an early bloom decay at the end of spring. Accordingly, the DMS concentrations are the lowest (Table S4) with an early spring rise corresponding to the

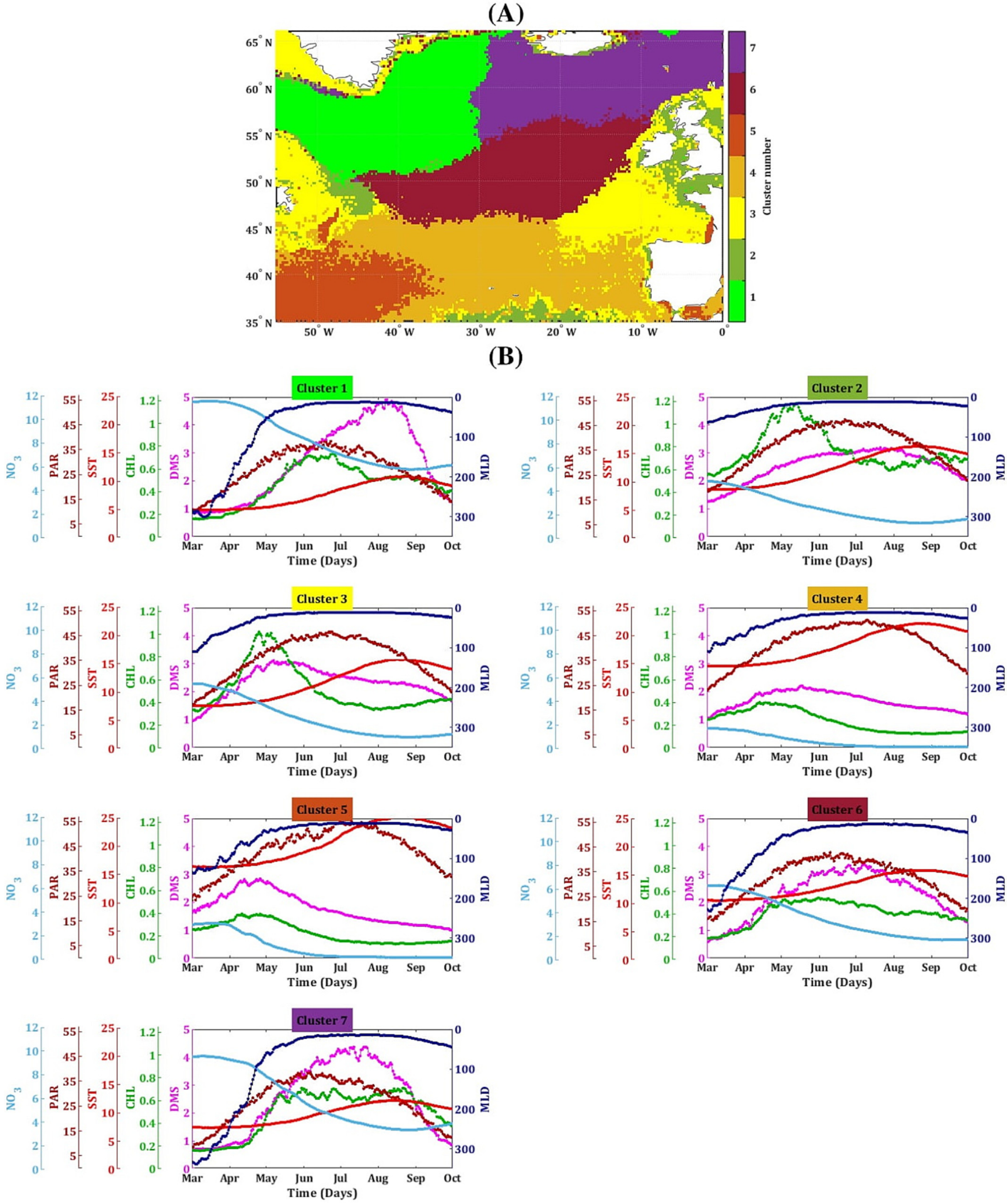


Fig. 4. (A) Spatial distribution of the clusters obtained from the k-means + + analysis of daily DMS (GPR) over the NA in 1998–2021. (B) mean DMS and the predictors annual cycles (1998–2021) in each cluster. Colors on y-axes distinguish between variables.

phytoplankton bloom. Clusters 2 and 3 are limited to coastal areas characterized by a smooth seasonal cycle associated with an early bloom (maximum biomass in early May).

Eventually, we observe an inverse seasonality between NO_3 and DMS demonstrating the complexity of DMS variability and showing the importance of other predictors (CHL and MLD) over nutrient concentration. The consumption of nutrients by phytoplankton when ambient conditions turn favorable (e.g., high SST and PAR in summer) and the summertime stratification, which suppresses nutrient upwelling, may explain the observed DMS-NO_3 reversed seasonality.

3.6. DMS flux distributions

The daily sea-to-air F_{DMS} are calculated using the Goddijn-Murphy et al. (2012) gas transfer velocity parameterization (see Section 2.6 for details)

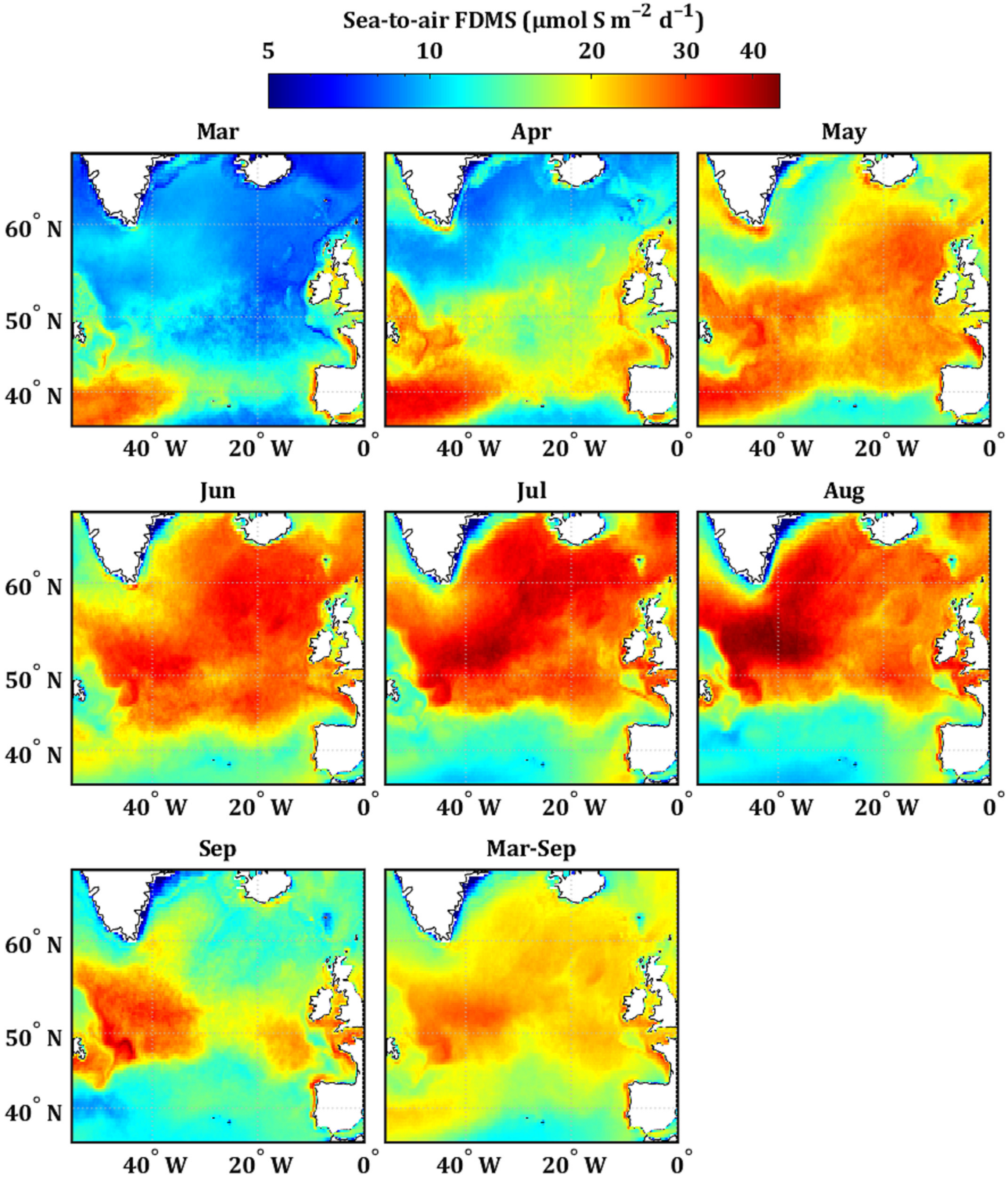


Fig. 5. Climatological DMS sea–air fluxes over 1998–2021 derived from the predicted DMS concentrations (GPR) and Goddijn-Murphy et al. (2012) parametrization.

Table 6
Monthly mean (±spatial standard deviation) of sea–air DMS fluxes in the North Atlantic region (35°–66° N, 55°–00°W). Total cumulative fluxes (integrated over the area of each pixel in the domain) of DMS-derived sulfur (Tg) calculated from the GPR and compared against G18 prediction and in situ L11 climatology. For L11, the flux calculation uses the climatic monthly mean sea surface temperature and wind speed during the study period.

	F_{DMS} (μmol S m^{-2} d^{-1})	Cumulative sulfur emissions (Tg)			Emission relative difference (%)	
	GPR	GPR	G18	L11	$\frac{GPR-G18}{G18}$	$\frac{GPR-L11}{L11}$
Mar	12.3 ± 4.9	0.30	0.24	0.36	+26	−17
Apr	16.8 ± 6.2	0.39	0.46	0.65	−16	−40
May	21.1 ± 5.0	0.49	0.75	0.97	−35	−50
Jun	22.9 ± 6.4	0.50	0.74	1.00	−33	−50
Jul	23.2 ± 8.6	0.51	0.63	0.93	−19	−45
Aug	22.2 ± 9.0	0.49	0.50	0.82	−02	−40
Sep	16.8 ± 5.3	0.37	0.36	0.37	+04	+01
Mar-Sep	19.3 ± 3.7	3.04	3.66	5.39	−17	−44

and the DMS derived from GPR predictions. The monthly climatology of F_{DMS} computed from daily (0.25° × 0.25°) data is presented in Fig. 5. Seasonal variation in F_{DMS} is evident and mostly related to the DMS cycle rather than wind speed. Predicted average F_{DMS} in the NA ranges from 12.3 ± 4.9 μmol m^{-2} d^{-1} in March to 23.2 ± 8.6 μmol m^{-2} d^{-1} in July, with an average of 19.3 ± 3.7 μmol m^{-2} d^{-1} from March to September (Table 6). In early spring (March–April), F_{DMS} above 50°N is lower than below (in particular, to the west), despite higher wind speeds (Fig. S8). Conversely, above 50°N F_{DMS} increases from May and peaks in summer (June–August), with the northern part of the domain characterized by a significantly higher DMS sea-to-air transfer than the southern region. The inverse seasonality of wind speed and F_{DMS} indicates that F_{DMS} is primarily driven by high seawater DMS concentration, and thus the gas transfer velocity (Eq. (15)) parameterized by wind speed appears to have a minor impact on F_{DMS} seasonality over the NA Ocean.

Furthermore, the total cumulative fluxes over the NA domain have been calculated, to obtain DMS-derived sulfur emissions (Tg) from the daily GPR/G18 and monthly L11 climatology. For L11, the flux calculation uses the climatological monthly mean sea surface temperature and wind speed during the study period from 1998 to 2021. L11 produces the highest spring-summer DMS flux (5.39 Tg from March to September) presented in Table 6. The GPR model yielded integrated sea–air DMS-derived sulfur emission of 3.04 Tg, which is 17 % and 44 % lower on average than G18 and L11 estimates.

4. Conclusions

In the present study, machine learning models were evaluated for the prediction of daily seawater DMS concentrations and their associated sea-to-air emission fluxes. For this aim, ship-based DMS observations were jointly analyzed with satellite oceanographic data and Copernicus-modelled data. The study domain corresponds to the NA ocean and the reconstructed dataset covers 24 years, from 1998 to 2021, at 0.25° × 0.25° spatial resolution. Our results show that the optimal exponential GPR (selected as the best performing machine learning model) captures up to 71 % of the observed NA DMS surface concentration, substantially improving the predictive strength over traditional empirical algorithms. The GPR resulted an efficient tool for obtaining reliable surface seawater DMS concentrations and it may be successful in other oceanic regions or over the entire global ocean as well. This new modeling approach predicts spatial DMS distributions that are coherent with the underlying patterns of oceanographic variability. Most notably, DMS concentrations over the NA, and the resulting sea-to-air fluxes, resulted strongly driven by regional and mesoscale patterns in phytoplankton biomass and seawater vertical mixing dynamics. This results in a notable latitudinal gradient of the DMS concentrations across the 50th parallel, particularly evident during the warm season, which previous algorithms fail to reconstruct.

The cluster analysis further reveals the marked spatial variability of DMS concentration in the NA, showing how varying seasonal patterns characterize different regions of the domain. The amplitude and extent of the DMS annual cycle are controlled by the joint impact of the predictors, most noticeably, phytoplankton biomass abundance and vertical mixing play intertwined roles in this. We observe that when there is a marked seasonal variation in phytoplankton biomass, also the predicted DMS concentrations show a remarkable seasonality, with sustained peaks lasting through spring and summer. Conversely, oceanic regions characterized by a relatively flat phytoplankton biomass seasonal trend or by short-lasting peaks are associated with lower summertime DMS concentrations. Seawater vertical mixing likely plays a double role in this. On the one hand, physical vertical mixing in the ocean strongly influences seawater productivity. Oceanic regions which undergo deep mixing during the winter season (e.g., the North region of the studied domain) are among the most productive areas as the deep mixing replenishes near-surface nutrients allowing for long-lasting phytoplankton blooming, once the right light conditions occur (i.e., spring). On the other hand, seawater vertical mixing may affect the duration and magnitude of the DMS concentration peaks also by stressing phytoplankton cells trapped in a shallow surface mixing layer, which favors the production and release of significant amounts of DMS in seawater and reduces its dilution.

The GPR-based sea-air flux in the North Atlantic is estimated to be 3.04 Tg S, from March to September, highlighting the importance of the study area as a globally significant sulfur source to the atmosphere. Improving the predictive accuracy of DMS concentration in seawater on both spatial and temporal scales allows for a better understanding of the mechanisms underlying DMS cycling and atmospheric export. More reliable DMS data will also aid in improving predictions of biogenic sulfur aerosol concentrations in the atmosphere potentially leading to a better understanding of the oceanic sulfur-aerosol-cloud interactions. Future studies on the interannual variability and long-term trends of DMS emissions are also possible thanks to the presented high-resolution dataset and will be an object of future investigations.

Funding

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

CRediT authorship contribution statement

Karam Mansour: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Stefano Decesari:** Funding acquisition, Project administration, Investigation, Writing – review & editing. **Darius Ceburnis:** Investigation, Writing – review & editing. **Jurgita Ovadnevaite:** Investigation, Writing – review & editing. **Matteo Rinaldi:** Conceptualization, Investigation, Resources, Writing – review & editing.

Data availability

The data sources used in the present study are listed in Table 1 (Main Manuscript). DMS and F_{DMS} monthly climatology from the GPR model can be accessed at https://doi.org/10.5281/zenodo.7030958. The codes used in this study are available based upon request from the corresponding author.

Declaration of competing interest

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 EU Copernicus Marine Environment Monitoring Service (CMEMS) for the provision of satellite ocean color data and modelled data, the Copernicus climate change service (C3S) for providing ERA5 reanalysis meteorological data, the ESA Climate Change Initiative and the SST CCI project for the provision of SST data, and the PMEL and NAAMES projects for the observed DMS measurements.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2023.162123.

References

Arthur, D., Vassilvitskii, S., SIAM/ACM, 2007. k-Means plus plus: the advantages of careful seeding. Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1027–1035.

Asante-Okyere, S., Shen, C.B., Ziggah, Y.Y., Rulegeya, M.M., Zhu, X.F., 2018. Investigating the predictive performance of Gaussian process regression in evaluating reservoir porosity and permeability. Energies 11.

Becagli, S., Lazzara, L., Fani, F., Marchese, C., Traversi, R., Severi, M., et al., 2013. Relationship between methanesulfonate (MS-) in atmospheric particulate and remotely sensed phytoplankton activity in oligo-mesotrophic central Mediterranean Sea. Atmos. Environ. 79, 681–688.

Behrenfeld, M.J., Moore, R.H., Hostetler, C.A., Graff, J., Gaube, P., Russell, L.M., et al., 2019. The North Atlantic Aerosol and Marine Ecosystem Study (NAAMES): science motive and mission overview. Front. Mar. Sci. 6.

Bell, T.G., Porter, J.G., Wang, W.L., Lawler, M.J., Boss, E., Behrenfeld, M.J., et al., 2021. Predictability of seawater DMS during the North Atlantic Aerosol and Marine Ecosystem Study (NAAMES). Front. Mar. Sci. 7.

Caliński, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Commun.Stat. 3, 1–27.

Charlson, R.J., Lovelock, J.E., Andreae, M.O., Warren, S.G., 1987. Oceanic phytoplankton, atmospheric sulfur,cloud albedo and climate. Nature 326, 655–661.

Diaz, B., Knowles, B., Johns, C.T., Laber, C.P., Bondoc, K.G.V., Haramaty, L., et al., 2021. Seasonal mixed layer depth shapes phytoplankton physiology, viral production, and accumulation in the North Atlantic. Nat. Commun. 12.

Dickson, D.M.J., Kirst, G.O., 1987. Osmotic adjustment in marine eukaryotic algae - the role of inorganic-ions, quaternary ammonium, tertiary sulfonium and carbohydrate solutes. 1. Diatoms and a rhodophyte. New Phytol. 106, 645–655.

Facchini, M.C., Decesari, S., Rinaldi, M., Carbone, C., Finessi, E., Mircea, M., et al., 2008. Important source of marine secondary organic aerosol from biogenic amines. Environ. Sci. Technol. 42, 9116–9121.

Friedland, K.D., Record, N.R., Asch, R.G., Kristiansen, T., Saba, V.S., Drinkwater, K.F., et al., 2016. Seasonal phytoplankton blooms in the North Atlantic linked to the overwintering strategies of copepods. Elementa 4, 1–19.

Gali, M., Devred, E., Levasseur, M., Royer, S.J., Babin, M., 2015. A remote sensing algorithm for planktonic dimethylsulfoniopropionate (DMSP) and an analysis of global patterns. Remote Sens. Environ. 171, 171–184.

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

Goddijn-Murphy, L., Woolf, D.K., Marandino, C., 2012. Space-based retrievals of air-sea gas transfer velocities using altimeters: calibration for dimethyl sulfide. J. Geophys. Res. Oceans 117.

Herr, A.E., Kiene, R.P., Dacey, J.W.H., Tortell, P.D., 2019. Patterns and drivers of dimethylsulfide concentration in the northeast subarctic Pacific across multiple spatial and temporal scales. Biogeosciences 16, 1729–1754.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horanyi, A., Munoz-Sabater, J., et al., 2020. The ERA5 global reanalysis. Q. J. R. Meteorol. Soc. 146, 1999–2049.

Hill, R.W., White, B.A., Cottrell, M.T., Dacey, J.W.H., 1998. Virus-mediated total release of dimethylsulfoniopropionate from marine phytoplankton: a potential climate process. Aquat. Microb. Ecol. 14, 1–6.

Kettle, A.J., Andreae, M.O., Amouroux, D., Andreae, T.W., Bates, T.S., Berresheim, H., et al., 1999. A global database of sea surface dimethylsulfide (DMS) measurements and a procedure to predict sea surface DMS as a function of latitude, longitude, and month. Glob. Biogeochem. Cycles 13, 399–444.

Kwint, R.L.J., Kramer, K.J.M., 1995. Dimethylsulfide production by plankton communities. Mar. Ecol. Prog. Ser. 121, 227–237.

Lacour, L., Claustre, H., Prieur, L., D'Ortenzio, F., 2015. Phytoplankton biomass cycles in the North Atlantic subpolar gyre: a similar mechanism for two different blooms in the Labrador Sea. Geophys. Res. Lett. 42, 5403–5410.

Lana, A., Bell, T.G., Simo, R., Vallina, S.M., Ballabrera-Poy, J., Kettle, A.J., et al., 2011. An updated climatology of surface dimethlysulfide concentrations and emission fluxes in the global ocean. Glob. Biogeochem. Cycles 25.

Laroche, D., Vezina, A.F., Levasseur, M., Gosselin, M., Stefels, J., Keller, M.D., et al., 1999. DMSP synthesis and exudation in phytoplankton: a modeling approach. Mar. Ecol. Prog. Ser. 180, 37–49.

Lloyd, S.P., 1982. Least-squares quantization in PCM. IEEE Trans. Inf. Theory 28, 129–137.

Mansour, K., Decesari, S., Bellacicco, M., Marullo, S., Santoleri, R., Bonasoni, P., et al., 2020a. Particulate methanesulfonic acid over the central Mediterranean Sea: source region identification and relationship with phytoplankton activity. Atmos. Res. 237.

Mansour, K., Decesari, S., Facchini, M.C., Belosi, F., Paglione, M., Sandrini, S., et al., 2020. Linking marine biological activity to aerosol chemical composition and cloud-relevant properties over the North Atlantic Ocean. J. Geophys. Res.-Atmos. 125.

Mansour, K., Rinaldi, M., Preißler, J., Decesari, S., Ovadnevaite, J., Ceburnis, D., et al., 2022. Phytoplankton impact on marine cloud microphysical properties over the Northeast Atlantic Ocean. J. Geophys. Res. Atmos. 127, e2021JD036355.

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

Merchant, C.J., Embury, O., Bulgin, C.E., Block, T., Corlett, G.K., Fiedler, E., et al., 2019. Satellite-based time-series of sea-surface temperature since 1981 for climate applications. Sci. Data 6.

Morel, A., Smith, R.C., 1974. Relation between total quanta and total energy for aquatic photosynthesis. Limnol. Oceanogr. 19, 591–600.

Morel, A., Huot, Y., Gentili, B., Werdell, P.J., Hooker, S.B., Franz, B.A., 2007. Examining the consistency of products derived from various ocean color sensors in open ocean (Case 1) waters in the perspective of a multi-sensor approach. Remote Sens. Environ. 111, 69–88.

O'Dowd, C.D., Facchini, M.C., Cavalli, F., Ceburnis, D., Mircea, M., Decesari, S., et al., 2004. Biogenically driven organic contribution to marine aerosol. Nature 431, 676–680.

Royer, S.J., Mahajan, A.S., Gali, M., Saltzman, E., Simo, R., 2015. Small-scale variability patterns of DMS and phytoplankton in surface waters of the tropical and subtropical Atlantic, Indian, and Pacific Oceans. Geophys. Res. Lett. 42, 475–483.

Royer, S.J., Gali, M., Mahajan, A.S., Ross, O.N., Perez, G.L., Saltzman, E.S., et al., 2016. A high-resolution time-depth view of dimethylsulphide cycling in the surface sea. Sci. Rep. 6.

Saltzman, E.S., King, D.B., Holmen, K., Leck, C., 1993. Experimental determination of the diffusion coefficient of dimethylsulfide in water. J. Geophys. Res. Oceans 98, 16481–16486.

Shi, C.M., Wei, B.T., Wei, S.L., Wang, W., Liu, H., Liu, J.L., 2021. A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. EURASIP J. Wirel. Commun. Netw. 2021.

Siegel, D.A., Doney, S.C., Yoder, J.A., 2002. The North Atlantic spring phytoplankton bloom and Sverdrup's critical depth hypothesis. Science 296, 730–733.

Simo, R., Dachs, J., 2002. Global ocean emission of dimethylsulfide predicted from biogeophysical data. Glob. Biogeochem. Cycles 16.

Sunda, W., Kieber, D.J., Kiene, R.P., Huntsman, S., 2002. An antioxidant function for DMSP and DMS in marine algae. Nature 418, 317–320.

Syakur, M.A., Khotimah, B.K., Rochman, E.M.S., Satoto, B.D., 2017. Integration K-means clustering method and elbow method for identification of the best customer profile cluster. 2nd International Conference on Vocational Education and Electrical Engineering (ICVEE). 336. Univ Negeri Surabaya, Fac Engn, Dept Elect Engn, Surabaya, INDONESIA.

Toole, D.A., Siegel, D.A., 2004. Light-driven cycling of dimethylsulfide (DMS) in the Sargasso Sea: closing the loop. Geophys.Res.Lett. 31.

Vallina, S.M., Simo, R., 2007. Strong relationship between DMS and the solar radiation dose over the global surface ocean. Science 315, 506–508.

Verrelst, J., Rivera, J.P., Gitelson, A., Delegido, J., Moreno, J., Camps-Valls, G., 2016. Spectral band selection for vegetation properties retrieval using Gaussian processes regression. Int. J. Appl. Earth Obs. Geoinf. 52, 554–567.

Wang, S.L., Elliott, S., Maltrud, M., Cameron-Smith, P., 2015. Influence of explicit Phaeocystis parameterizations on the global distribution of marine dimethyl sulfide. J. Geophys. Res. Biogeosci. 120, 2158–2177.

Wang, W.L., Song, G.S., Primeau, F., Saltzman, E.S., Bell, T.G., Moore, J.K., 2020. Global ocean dimethyl sulfide climatology estimated from observations and an artificial neural network. Biogeosciences 17, 5335–5354.

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

Wolfe, G.V., Steinke, M., 1996. Grazing-activated production of dimethyl sulfide (DMS) by two clones of Emiliania huxleyi. Limnol. Oceanogr. 41, 1151–1160.

Zhuang, G.C., Yang, G.P., Yu, J.A., Gao, Y.A., 2011. Production of DMS and DMSP in different physiological stages and salinity conditions in two marine algae. Chin. J. Oceanol. Limnol. 29, 369–377.

Zindler, C., Marandino, C.A., Bange, H.W., Schutte, F., Saltzman, E.S., 2014. Nutrient availability determines dimethyl sulfide and isoprene distribution in the eastern Atlantic Ocean. Geophys. Res. Lett. 41, 3181–3188.