npj | climate and atmospheric science

Article

Published in partnership with CECCR at King Abdulaziz University

https://doi.org/10.1038/s41612-024-00830-y

Nested cross-validation Gaussian process to model dimethylsulfide mesoscale variations in warm oligotrophic Mediterranean seawater

Karam Mansour [1,2] ✉, Stefano Decesari [1], Marco Paglione [1], Silvia Becagli [3,4] & Matteo Rinaldi [1]

The study proposes an approach to elucidate spatiotemporal mesoscale variations of seawater Dimethylsulfide (DMS) concentrations, the largest natural source of atmospheric sulfur aerosol, based on the Gaussian Process Regression (GPR) machine learning model. Presently, the GPR was trained and evaluated by nested cross-validation across the warm-oligotrophic Mediterranean Sea, a climate hot spot region, leveraging the high-resolution satellite measurements and Mediterranean physical reanalysis together with in-situ DMS observations. The end product is daily gridded fields with a spatial resolution of 0.083° × 0.083° (~9 km) that spans 23 years (1998–2020). Extensive observations of atmospheric methanesulfonic acid (MSA), a typical biogenic secondary aerosol component from DMS oxidation, are consistent with the parameterized high-resolution estimates of sea-to-air DMS flux (F_{DMS}). This represents substantial progress over existing coarse-resolution DMS global maps which do not accurately depict the seasonal patterns of MSA in the Mediterranean atmospheric boundary layer.

Accurate estimation of sea-to-air dimethylsulfide flux (F_{DMS}), the largest natural source of sulfur aerosol particles in the atmosphere, requires a reliable prediction of seawater dimethylsulfide (DMS) concentration. DMS is a volatile biogenic gas produced by marine microorganisms that, when released into the atmosphere, influences the formation of aerosols and affects cloud properties[1–3] and consequently planetary albedo and climate[4]. Many attempts have been made to simulate the global DMS distributions, the first of which was introduced in 1999 by interpolating existing surface observations to generate monthly maps of 1° × 1° spatial resolution[5]. Twelve years later, improved monthly DMS global maps were created using about three times the number of data points as the initial version[6]. For a decade, most atmospheric investigations used these maps[6] as a reference product for DMS emissions. Recently, the DMS climatology maps have been updated once more[7], referred to as DMS-Rev3, with a roughly 18-fold increase in the raw database as compared to the second version[6] which results in more realistic monthly DMS estimates than the previous releases. Concurrently, empirical algorithms were developed to model DMS distributions as a function of controlling parameters like the ratios of chlorophyll-a concentration (CHL) to mixed layer depth (MLD)[8], solar radiation dose[9], photosynthetically active radiation (PAR), and satellite-based Dimethylsulfoniopropionate (DMSP) concentrations[10,11], the major precursor of marine DMS. Detailed information on the existing DMS retrieval algorithms can be found in ref. 12.

Machine learning (ML) algorithms have been applied to create climatological DMS fields that have the advantage of capturing nonlinear interactions between environmental variables and DMS concentrations. Wang et al.[13] employed artificial neural networks (ANN) for constructing global monthly climatology of DMS maps at 1° × 1° spatial resolution, hereafter referred to as W20. Although global DMS products are believed to be acceptable for qualitatively describing seasonal variations in DMS[7,13], their predictive ability tends to diminish at regional scales, failing to precisely resolve the oceanic mesoscale and sub-mesoscale spatial patterns or shorter temporal scale variability[12,14–17]. As a result, there is significant uncertainty in the regional distributions of DMS concentration and fluxes amongst various global products[18] owing to the limited space resolution.

Regionally, the monthly DMS climatology in the northeast Pacific Ocean was reconstructed at a higher spatial resolution of 0.25° × 0.25° using random forest regression and ANN[17]. The study concluded that DMS

[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]Department of Chemistry "Ugo Schiff", University of Florence, Sesto F. no (FI), Florence, 50019, Italy. [4]Institute of Polar Sciences, National Research Council (CNR-ISP), Venice Mestre, 30172, Italy. ✉e-mail: k.mansour@isac.cnr.it

patterns are linked to mesoscale oceanic variability; such patterns would have remained obscure in the coarser resolution products[12]. In terms of temporal coverage, climatology maps cannot properly analyze present and future DMS emission trends under global climate change scenarios[12,19]. Addressing this issue, Mansour et al.[12] generated the first grided daily time series of DMS concentrations and associated F_{DMS} at a spatial resolution of 0.25° × 0.25°. The dataset was generated for the North Atlantic using the GPR model. The DMS time series made it possible to elucidate high-frequency spatial and temporal patterns in DMS variability, proving the usefulness of a cutting-edge technique (GPR) for predicting sea surface DMS concentration and flux. Finding trustworthy high-resolution DMS concentration data is essential since changes in DMS dynamics usually take place on the days-to-week timescales associated with meteorological forcing[16]. Another advantage of the improved derived sea-to-air F_{DMS} time series is that it can be utilized to predict and comprehend the dynamics of marine-derived biogenic sulfur aerosol concentrations and their radiative effects[20]. Furthermore, carrying out relevant scientific studies aiming at reducing aerosol-cloud interaction uncertainty in climate models requires long-term, continuous, and high spatiotemporal resolution datasets.

The Mediterranean (MED) Sea, the object of this study, represents one of the most complex marine ecosystems on the planet[21], with varied physical and chemical processes occurring at different space-time scales such as deep-water formation, thermohaline circulation, and subbasin gyres. MED has been categorized as an oligotrophic basin[22] due to the typically low primary production, with notable bio-regionalization in phytoplankton biomass and abundance at the subbasin to regional scales[23,24]. Although biogenic sulfate aerosol contributes significantly to the sulfur burden in the MED atmosphere (estimated at more than 26%)[25–27] which is contributed by extensive urban pollution sources surrounding the basin as well as from ship traffic, scarce information is available on the spatio-temporal distribution of seawater DMS concentration and, particularly, sea-to-air flux. Previous findings based on short-term observations revealed that the MED sea, like most of the oligotrophic basins, presents the so-called summer DMS paradox[28,29] (elevated summer DMS concentrations coupled to low surface CHL levels). DMS production in the MED sea is irradiance-dependent[30,31] and is not proportional to total phytoplankton biomass[32,33]. Essentially, the DMS concentrations peak 1-2 months after CHL[33], following phytoplankton succession[34,35] in stressed cells or grazing-derived production[36,37], as well as physiological adjustments[31,38].

The study aims to assess the plausibility of applying the GPR model, which has been successfully applied in the North Atlantic ocean[12], to predict long-term (1998-2020) high-resolution gridded fields of daily DMS and F_{DMS} time-series covering the MED domain at 0.083° × 0.083° (grid cell area between 60 to 74 km^2) spatial resolution. Moreover, it evaluates how well the nested cross-validation performs while evaluating ML models with few observations. A comparison with currently available DMS climatology products has been conducted to show the uncertainty level of their distributions over the regional seas. We explain the seasonal patterns of the predicted DMS and F_{DMS} and how they follow the long-term measurements of MSA, a biogenic aerosol component of DMS oxidation. After evaluating the most popular ML regression models (Table 1 of ref. 20), GPR was shown to be the best-performing model for reconstructing DMS observations. The model was built employing the available in-situ sea surface DMS concentrations[5,16] (Fig. S1), high-resolution satellite observations of CHL, sea surface temperature (SST), and PAR as well as sea surface salinity (SSS) and MLD from the novel MED physical reanalysis[39]. To gain a better understanding of the links between these parameters and both DMS and F_{DMS} at different time scales, the spatio-temporal variations of these parameters were investigated over 23 years by performing the empirical orthogonal function (EOF) analysis[40] on the daily gridded datasets.

Results

GPR evaluation and data generation

Using the 5-fold nested cross-validation (nCV) approach (see Methods), we assess the GPR model on test subsets of the outer loop (Fig. 1a). The aim is to minimize overfitting and ensure that the trained/cross-validated model can be adapted to another dataset impartially. Figure 1b presents the scatter plot between observed and predicted seawater DMS concentrations by GPR. The average evaluation measures, on the 5 test folds, exhibit that GPR achieves a coefficient of determination (R^2) of 0.71 and root mean square error (RMSE) of 0.12. Notably, the model performance is impartial across the various inner cross-validation subsets, with an R^2 ranging from 0.63 to 0.72 and an RMSE between 0.12 and 0.14 (Fig. S2). The performance is consistent with the results given by the GPR over the North Atlantic domain ($R^2 = 0.71$ and RMSE = 0.21; on the test fold)[12], by implementing nearly four times as many data points in building the model. Contextually, previous studies reported similar prediction performances by using ANN to estimate the monthly climatology of global ocean DMS distributions ($R^2 = 0.66$)[13] and to characterize spatio-temporal DMS variability in the Yellow and East China Sea ($R^2 = 0.71$)[41]. The random forest regression and ANN caught up to 62% of the observed monthly climatology of seawater DMS fluctuation in the northeast subarctic Pacific[17].

We compared the GPR model as a tool to predict seawater DMS concentrations with the previously published empirical methods[8–10] and the ANN[13] of W20. We adapted an ANN model with the same hyperparameters as W20 (*i.e.*, one input layer, two hidden layers, one output layer, and a regulation parameter value of 0.001) keeping the same potential predictors used in GPR, for proper comparison. The results (Fig. 1c) show that GPR has much higher prediction accuracy with a markedly higher R^2 value and lower mean absolute error (MAE). GPR achieves MAE of 0.52 μmol m^{-3}, which is 59% lower than the most skilled empirical method[8], MAE = 1.26 μmol m^{-3}, and 7% lower than ANN, MAE = 0.56 μmol m^{-3}. This indicates a major improvement in the representation of the regional seawater DMS variability in the MED sea by GPR. The trained GPR model was used to calculate daily seawater DMS concentrations (Fig. 2a), as well as sea-to-air F_{DMS} (Fig. 3a), at each pixel of the MED domain. The gridded data has a high resolution of 0.083° × 0.083° covering from 1998 to 2020. As an illustration of the data product, Fig. 3b presents a daily time series of DMS and F_{DMS} averaged over the entire MED domain.

The annual mean climatological distribution of DMS (Fig. 2a) across the 23-year study period shows that GPR-DMS values vary from 1.68 to 4.39 μmol m^{-3} (minimum and maximum over all the grids), with median value equal to 3.18 μmol m^{-3}. In general, DMS concentrations are higher in the eastern Mediterranean (EMED) than in the western Mediterranean (WMED). The south-central Mediterranean (CMED), front of Libyan coast, and the northeast Levantine basin have the highest DMS concentrations. The north Adriatic (at the mouth of the PO river) and north Aegean (near the Sea of Marmara) have the lowest DMS concentrations, which could be attributed to fresh- and low- saline water discharge into the MED. Laboratory-controlled experiments showed that the DMS accumulation rate is reduced as salinity decreases[42–45]. The Alboran subbasin, which is near the Gibraltar Strait and is affected by the input of Atlantic water, is another area displaying low DMS concentrations. In comparison to recent estimates of annual climatology from global products (Rev3 and W20)[7,13], it is noticed that their output failed to resolve DMS spatial variations at regional scales, as expected due to their coarser resolution. The gradual increase in DMS from WMED to EMED can be somewhat represented by the W20, but Rev3 regarded the MED domain as a bulk region where DMS presents reduced spatial variability or almost constant concentrations (Fig. 2 and Fig. S3). Concerning the average concentrations across the MED, the GPR-DMS has an annual mean of 3.14 ± 0.32 μmol m^{-3}, which is comparable with Rev3 (3.61 ± 0.15 μmol m^{-3}) and twice as high as W20 (1.75 ± 0.19 μmol m^{-3}). On an annual basis, GPR shows an approximate 13% decrease in DMS concentration compared to Rev3 (Fig. 2d) but is 81% greater than W20 (Fig. 2e), throughout the whole domain.

The MED sea-to-air F_{DMS} (Fig. 3a) is calculated to be 4.6 ± 1.1 μmol m^{-2} d^{-1} on an annual basis; similarly to the DMS distribution, significant regional differences have been observed, primarily the contrasting emissions between the WMED (relatively low emissions) and EMED (high emission rates). The integrated F_{DMS}-derived sulfur emissions over the whole domain

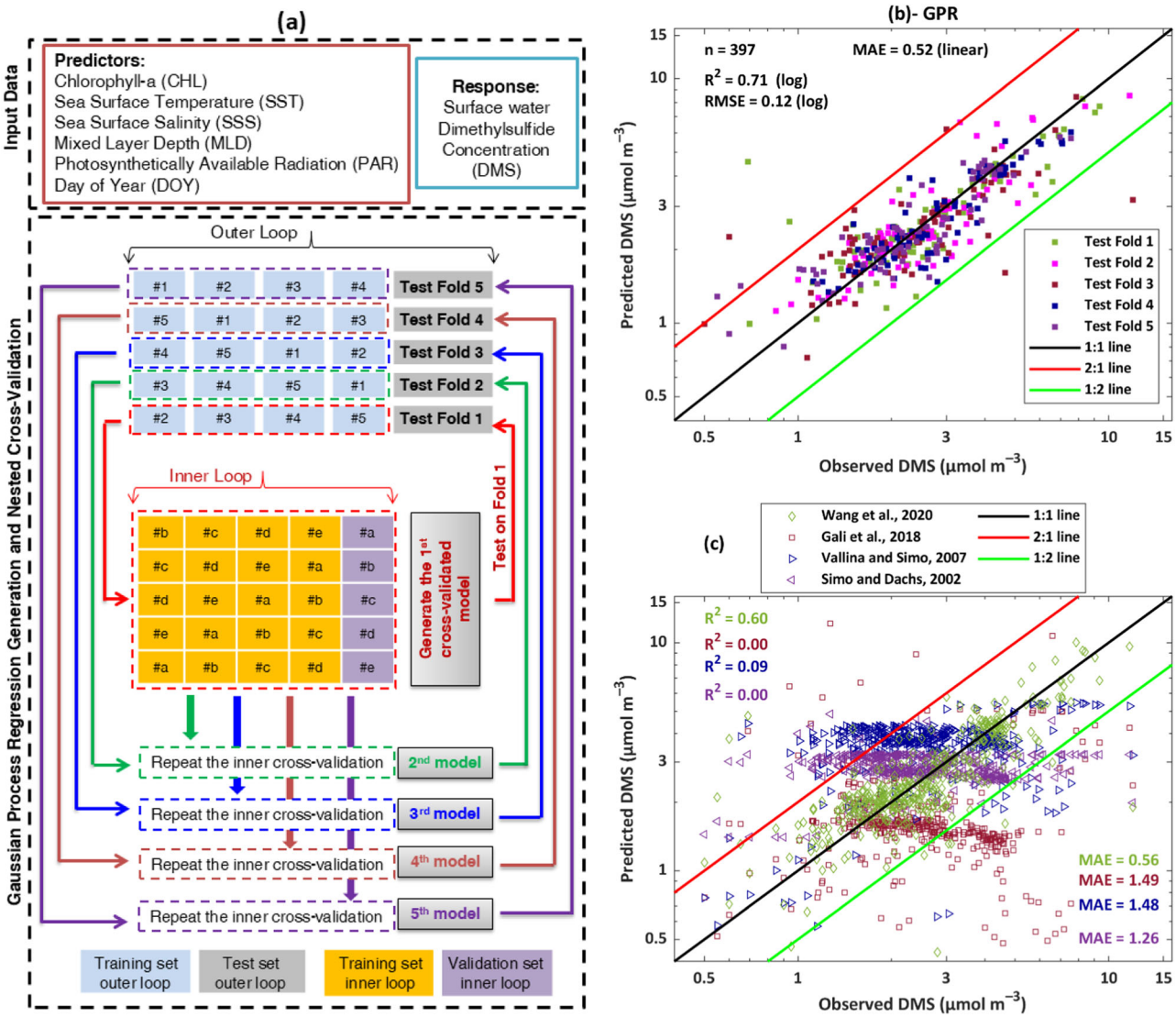

Fig. 1 | The GPR training and validation process. a The nested-cross validation strategy used to evaluate GPR. **b** comparison of predicted and observed seawater DMS by GPR, evaluated on the test folds of the outer loop. **c** comparison between estimated and observed DMS from the previous empirical algorithms and ANN.

have an average of 265.6 ± 7.6 Gg yr^{-1}, displaying notable interannual variability (Fig. 3b), with the largest flux recorded in 2016 (279.9 Gg) and the minimum flux occurred in 2004 (250.4 Gg). Such an average integrated flux is approximately 12% lower and 58% greater than Rev3 (301.6 Gg yr^{-1}) and W20 (168.4 Gg yr^{-1}) climatologies, respectively. It should be mentioned that the estimates of F_{DMS} from the aforementioned global products were computed using the same GPR-F_{DMS} parameterization method[46] because different approaches produced different results (see Discussion). This was done by using the monthly DMS concentrations (Rev3 and W20) and climatic monthly of SST and WS. This is required to conduct a proper comparison of GPR-F_{DMS} and other estimates, throughout the present study.

Monthly DMS and F_{DMS} distributions

In this section, we investigate the monthly climatology of DMS and F_{DMS} from GPR and compare them with the Rev3 and W20 products. The maps are presented in Fig. 4 and Fig. S4. With spring being a season of growth and autumn being a season of decay, the monthly mean sea surface DMS concentrations averaged over the years 1998–2020 (Fig. 4a) exhibit an evident seasonality minimum in winter (Dec to Feb) and maximum in summer (Jun–Jul). The EMED is characterized throughout the year by rather higher values than the WMED. It is observed that the spatial features of the DMS concentration in the oligotrophic MED sea do not match the distribution of CHL, a tracer of phytoplankton activity (Fig. S5) in agreement with previously reported observations[32,47]. We found that W20 underestimates the DMS concentration, particularly in spring and summer, when compared to GPR. Rev3 quantitatively displays very comparable DMS concentrations, however there is high overestimation in May and inconsistent low values in July (Fig. S4).

GPR-F_{DMS} monthly maps (Fig. 4b) show that the main seasonal cycle of DMS flux depends strongly on the seawater concentrations; GPR-F_{DMS} begin to rise in May and peaks in July, followed by a steady decline in September. December and January have the lowest GPR-F_{DMS} at roughly 2.1 ± 0.6 μmol m^{-2} d^{-1}, while July has the largest emission rate at 8.3 ± 2.0 μmol m^{-2} d^{-1}. According to the Rev3-F_{DMS} monthly maps (Fig. 4c), F_{DMS} peaks in May (9.3 ± 2.5 μmol m^{-2} d^{-1}), two months before GPR, with an incidental dip in July, followed by an increase in August; nonetheless, the minimum levels occur in January (1.3 ± 0.4 μmol m^{-2} d^{-1}). The W20-F_{DMS} (Fig. 4d), on the other hand, follows a regular seasonal cycle, with the lowest emission rate in January (0.8 ± 0.3 μmol m^{-2} d^{-1}) and the highest in August (3.8 ± 1.6 μmol m^{-2} d^{-1}). The differences (Rev3-GPR) in F_{DMS} maps (Fig. 4e) show irregular positive and negative values, with positive differences peaking in May (3.3 μmol m^{-2} d^{-1}) and negative differences peaking in July (−2.0 μmol m^{-2} d^{-1}). On the contrary, (W20-GPR) F_{DMS} climatology

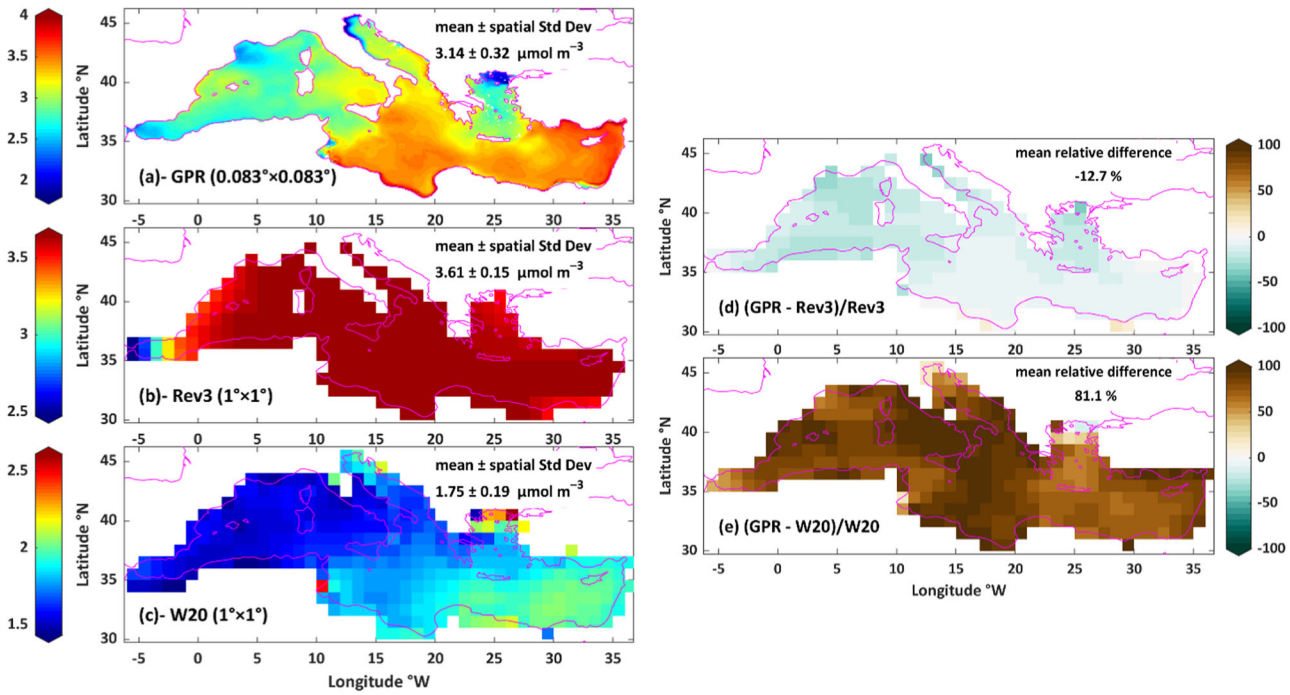

Fig. 2 | Annual climatology of seawater DMS concentrations. a spatial distribution of DMS based on the GPR model at 0.083° × 0.083° over 1998–2020. Climatology output at 1° × 1° based on Rev3 (**b**) and W20 (**c**) estimates. Note that the color scale is different for each product. The relative difference of GPR from Rev3 (**d**) and from W20 (**e**) as percentages.

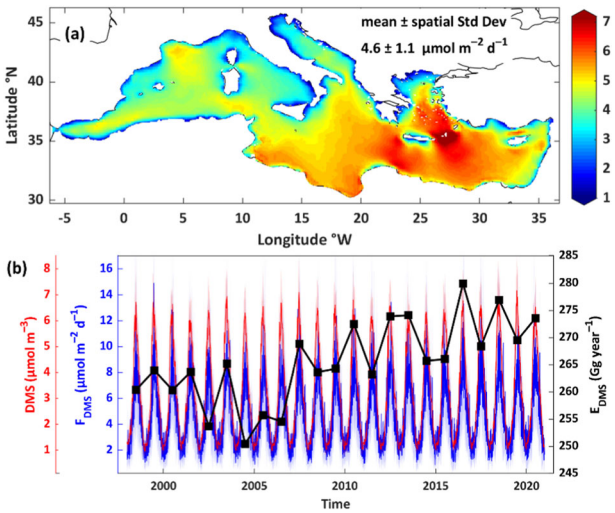

Fig. 3 | The sea-to-air F_{DMS} annual climatology and time series. a Annual distribution of F_{DMS} at 0.083° × 0.083° over 1998–2020 derived from the predicted DMS concentrations (GPR) and Blomquist et al.[46] parametrization. **b** daily time series of average DMS and F_{DMS} in the entire Mediterranean obtained by GPR in 1998–2020, whereas shaded areas represent ± spatial standard deviations. The black line represents the annual total cumulative emission fluxes (integrated over the area of each pixel in the domain) in Giga grams per year.

(Fig. 4d) shows lower values throughout the year, with the greatest discrepancies observed from May to July (up to −6 µmol m^{-2} d^{-1} in certain areas of the MED domain) as displayed in Fig. 4f.

These findings highlight the discrepancies between global products when estimating DMS emission flux on the regional scale. One shortcoming of both products (Rev3 and W20) is the horizontal resolution (1° × 1°), which pre-supposes that DMS remains constant across each grid area of the domain (between 8.8 × 10^3 and 10.7 × 10^3 km^2, in the case of the MED sea). This may be inconsistent with the MED sea's subbasin scale, mesoscale gyres, semi-enclosed nature, complex morphology and coastlines, and a wide range of physical and chemical processes that govern its productivity[23,48]. In terms of quantitative estimation of the DMS-derived sulfur aerosol budget, while the Rev3 product misses a large portion of grids over the Adriatic Sea, the W20 output provides DMS values over the majority of Italy's land area (Fig. 2), indicating a major problem with data quality at the regional scale.

EOF and drivers of DMS and F_{DMS} variabilities

The preceding discussions show that DMS and F_{DMS}, retrieved by GPR, display a wide spatiotemporal variability in the MED sea. Herein, the EOF analysis is applied to the gridded high-resolution dataset to investigate the DMS and F_{DMS} spatial variability patterns and how they change with time. Then, to better understand the probable interactions between them and their independent controllers at different space-time scales. The EOF analysis is one of the most extensively used methods for understanding spatio-temporal variability in oceanic and atmospheric data[40,49]. It allows us to identify the dominant modes of variability by decomposing the dataset into spatial modes (EOF modes that show the patterns of variability) and their associated time series or principal components (PCs), which quantifies the importance of each mode. The findings of the first three EOF modes, where the majority of the variance is explained, for DMS and F_{DMS}, as well as the governing parameters (CHL, MLD, SST, PAR, SSS, WS, and K_{DMS}), were evaluated in the present study. We applied the EOF analysis to the original daily time series during 1998–2020, in order to account for the main cycles of the studied parameters.

The spatial patterns of EOF modes and the normalized climatology (during 1998–2020) of their amplitude time series (PCs) for DMS and F_{DMS} are displayed in Fig. 5a–d. The original PCs time series are presented in Fig. S6. The first three EOF modes of the DMS (F_{DMS}) account for about 93.7% (64.8%) of the overall variance of the data, with the first mode (EOF1) alone explaining about 90% (50%). In parallel, the EOF1 of CHL, MLD, SST, PAR, SSS, WS, and K_{DMS} accounts for 56.4%, 77.8%, 95.8%, 83.7%, 22.4%, 33.2% and 29.2% of the variance, respectively (Figs. S7–S9).

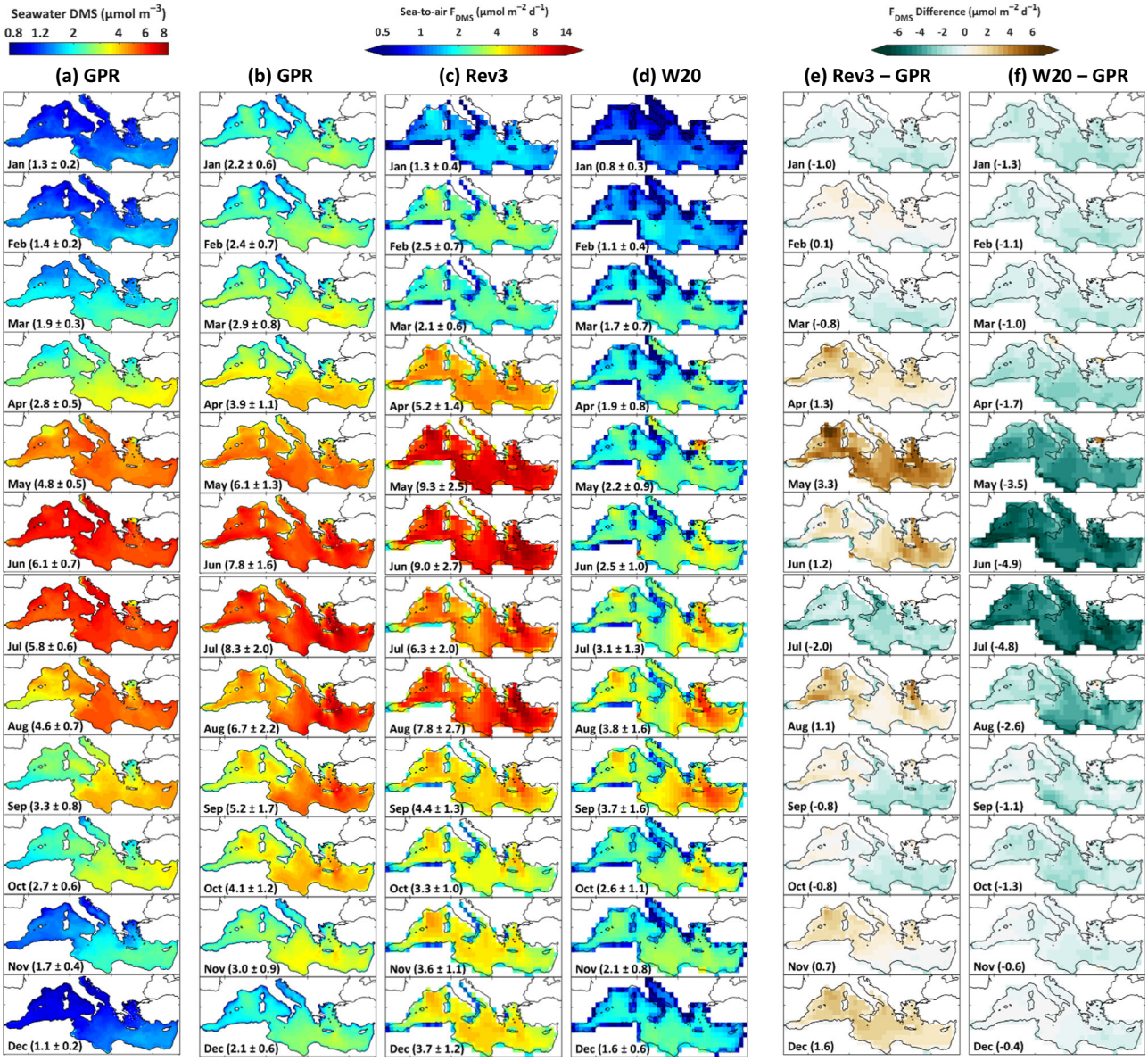

Fig. 4 | Monthly spatial distributions of sea surface DMS concentrations and related emission flux. a DMS maps based on GPR during 1998–2020, (**b**–**d**) sea-to-air F_{DMS} based on the Blomquist et al.[46] parametrization and the predicted DMS from GPR, Rev3 and W20. **e** the difference between Rev3-based F_{DMS} and GPR-based F_{DMS}. **f** the difference between the W20-based F_{DMS} and GPR-based F_{DMS}. The pixel resolution of (**a**) and (**b**) is 0.083° × 0.083° while (**c**–**f**) have 1° × 1° spatial resolution. The monthly (average ± spatial standard deviation) is shown in brackets in each panel.

The spatial patterns of EOF1 for DMS (Fig. 5a) and F_{DMS} (Fig. 5b) are positive throughout the MED domain, indicating an in-phase oscillation of the entire basin around the steady-state mean. This indicates the cycle crest emerges when the PC time series is positive too, and vice versa. The Tyrrhenian Sea and the north-central part of the basin including the Adriatic have the largest DMS annual amplitude, whereas the Alboran Sea, South Levantine, and the northernmost part of the Aegean Sea have the lowest variability. Besides, F_{DMS} displays the highest variability in the northwest part of the basin and the Aegean Sea, to the east of Crete Island. The PC1 time series (Fig. S6) and their normalized climatology during 1998-2020 of DMS (Fig. 5c) and F_{DMS} (Fig. 5d) show a clear annual cycle that is the leading component of variabilities. The subbasins showing higher EOF1 values are therefore characterized by a seasonal DMS cycle of greater amplitude. The maximum temporal amplitude is reached in June and July and the minimum in winter (Dec-Feb). The annual cycle of DMS is driven mainly by the available radiation for photosynthesis ($r = 0.93$; $p < 0.05$ between PC1-DMS and PC1-PAR) (Fig. 5e and Fig. S10). When low CHL and a shallow mixed layer coexist with increased PAR, it suggests that high DMS concentrations are linked to stressed phytoplankton cells stuck in a shallow surface mixing layer due to oxidative stress brought on by irradiance. Contextually, F_{DMS} seasonality is mainly governed by the seawater DMS concentrations ($r = 0.89$; $p < 0.05$; Fig. S11).

The DMS EOF2 mode accounts for around 3.1% of the overall variance. Its spatial pattern is a dipole, with opposite fluctuation between the EMED and WMED (plus Adriatic) subbasins. According to the DMS PC2 climatology, there are two peaks: a point of maximum by the end of April and a peak from end of August to start of September; there are also two troughs: a sharp minimum by the end of June and a broader minimum from January to February. Consequently, high-frequency variations in DMS peaked primarily in late spring and late summer in the EMED, simultaneously, the Adriatic and WMED seas show minimum amplitudes. The peak in late spring may correspond to phytoplankton succession occurring 1–2 months after the late winter and early spring bloom[34,35]. A possible interpretation of the second peak is that warm EMED water masses may

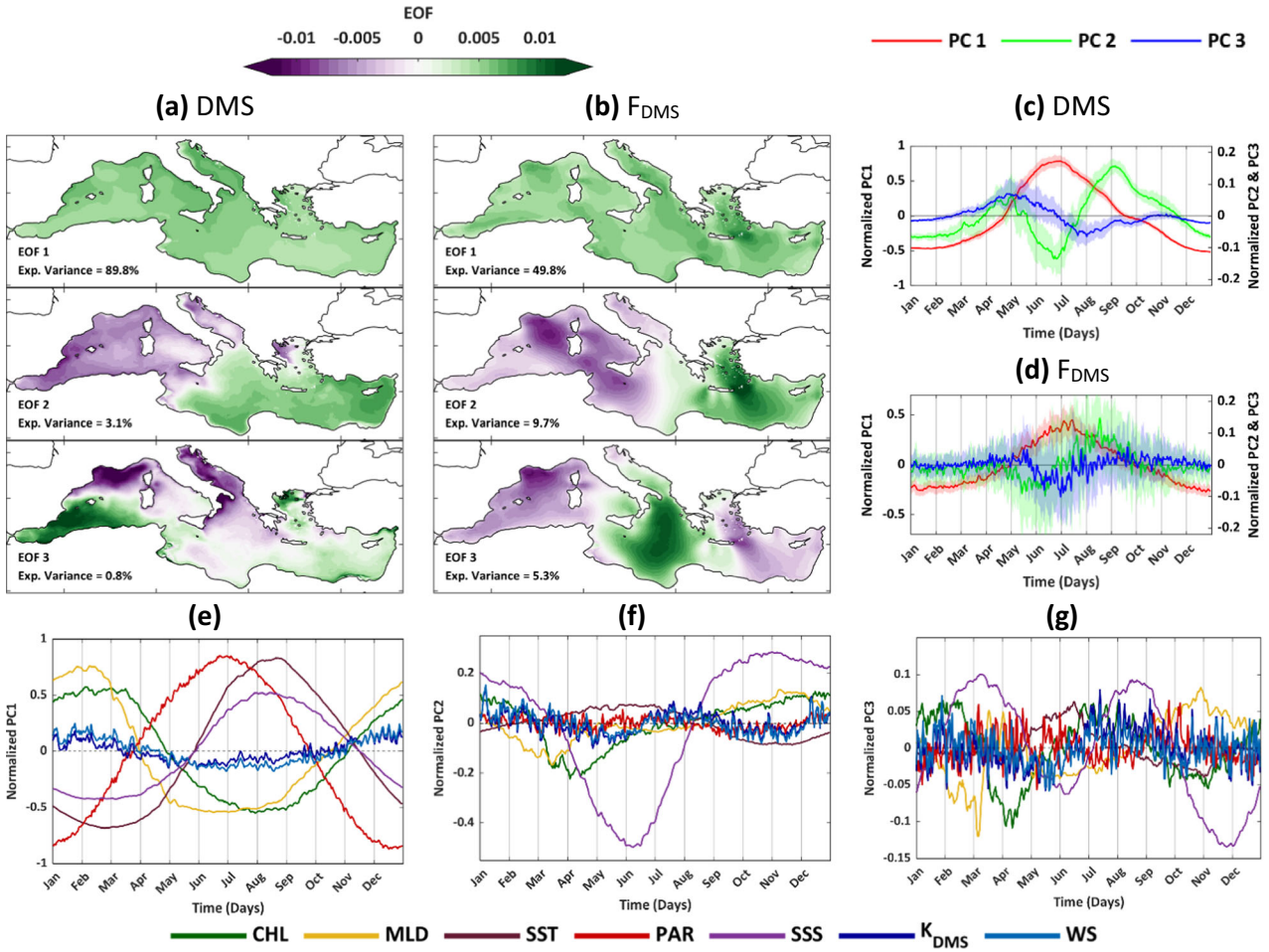

Fig. 5 | EOF analysis. The first three modes of EOF spatial patterns of (**a**) DMS and (**b**) F_{DMS} calculated from daily (0.083° × 0.083°) datasets generated by GPR and their corresponding daily climatology of PCs time series for the 1998–2020 period (**c**, **d**). **e**, **f**, and **g**, daily climatology of PCs time series of the environmental parameters controlling DMS and F_{DMS}. The PCs of each parameter have normalized from −1 to 1. Shaded areas represent ± standard deviations.

induce DMS production driven by stratification-stressed effects at high surface irradiance. This can be evidenced by the consistency between DMS-EOF2 and SST-EOF2 (Fig. S8) spatial patterns as well as the considerable association between their PCs time series (Fig. S10; $r = -0.37$; $p < 0.05$, being the negative sign due to the opposite spatial modes). In parallel, the EOF2 of F_{DMS} explains roughly 9.7% of the variance and divides the MED Sea into two regions with opposing phases. On one side, the central part along the Sicily channel, the Tyrrhenian subbasin, and the northwest section of the MED show relatively high F_{DMS} in May-June. The Levantine and Aegean subbasins, on the other side, exhibit an increase in emissions during August.

The DMS EOF3 mode accounts for a minor percentage of the variance (0.8%) and will not be discussed. Conversely, the F_{DMS} EOF3 accounts for about 5.3% of the total variance and its spatial distribution shows an out-of-phase oscillation with opposite peaks between the central part of the basin and the westernmost and easternmost parts of the basin. The maximum variability (with opposed phases) is observed in the Gulf of Lion and the central part of the basin. The variations of PC2 and PC3 of the F_{DMS} time series (Fig. S6 and Fig. 5d) are characterized by high-frequency oscillations (low periods; of the order of days). Such small-time scale variations of F_{DMS} can be driven by changes in wind speed and consequently, the DMS gas transfer velocity (K_{DMS}). This can be seen clearly from the positive correlation between PC2-F_{DMS} and both PC2-WS ($r = 0.67$; $p < 0.05$) and PC2-K_{DMS} ($r = 0.70$; $p < 0.05$) as well as a similar relationship between the PC3 time series of these components (Fig. S11). It is worthwhile to point out that the Gulf of Lion is vulnerable to strong short-term wind events (e.g., the Mistral wind[50]) which contribute to enhanced air-sea interactions[51].

In summary, the EOF analysis shows that the annual fluctuation is the leading component of DMS variability over the MED sea, captured by EOF1; such a component is driven mainly by physical conditions like the available solar radiation. Variations on a smaller time scale contribute to a minor part of the variability and occur with opposite phases in the western and eastern parts of the domain.

Also, the variability of F_{DMS} is mostly driven by the yearly cycle, even though the magnitude of the variation is lower than that of DMS concentration whilst EOF 2 and 3 are relatively more important for F_{DMS} than for DMS. The main driver of the yearly fluctuations in F_{DMS} is the seawater DMS concentration. Local winds seem to play an important role in the F_{DMS} short-term scale fluctuations over the MED sea, as highlighted in EOF2 and EOF3. Supporting that, the F_{DMS} monthly distributions (Fig. 4b) show a hotspot of high emission rates over the Aegean Sea (an area of low DMS concentrations) during summer (the season of low wind). This could be due to the impact of the intensified Etesian winds blowing over the Aegean Sea during summer[52]. This can be observed in the monthly fields of wind climatology (1998-2020) presented in Fig. S12.

F_{DMS} and MSA relationship: potential perspectives

The study's outcomes provide an inventory of marine biogenic DMS emissions that can be used to improve the modeling of biogenic sulfur aerosol concentrations[20] in the MED atmosphere. To investigate this possibility, we compare high-resolution constructed GPR-F_{DMS} with atmospheric particulate methanesulfonic acid (MSA) concentrations in background marine conditions. The MSA measurements were carried out at

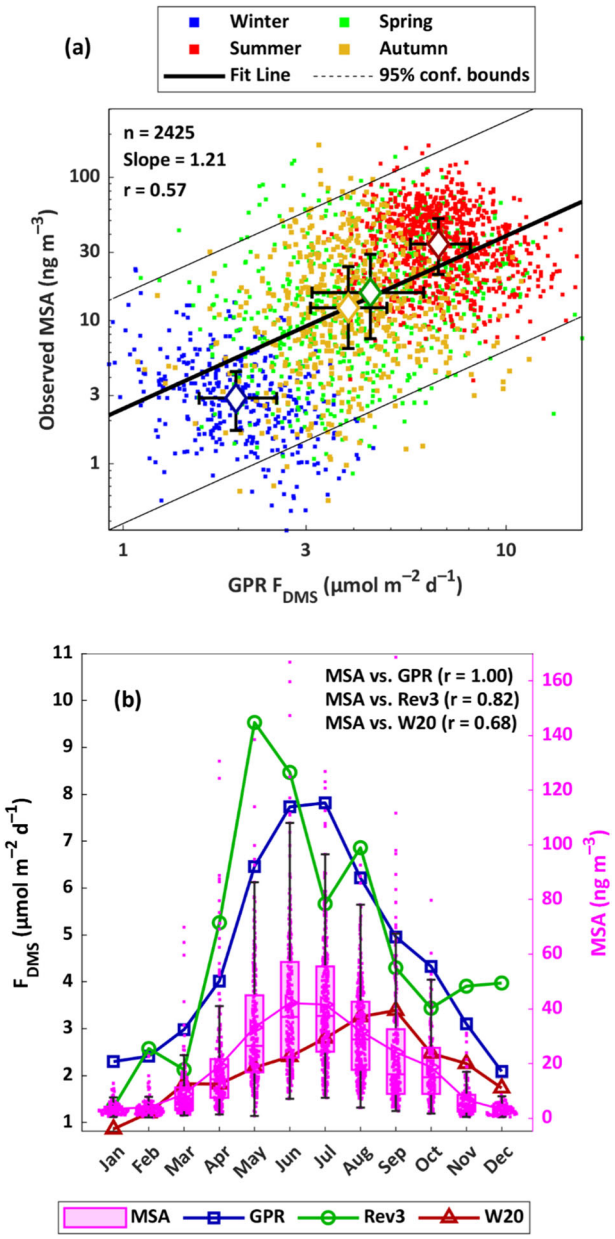

Fig. 6 | Relationship between F_{DMS} and MSA. a Scatter plot between MSA observed at Lampedusa measuring site (black filled square in Fig. S1) and the GPR F_{DMS} averaged inside the box delineated in Fig. S1. The F_{DMS} is calculated from the predicted DMS concentrations by GPR employing Blomquist et al.[46] parametrization. The diamond symbols represent the seasonal median and the whiskers extend to the 1st and 3rd quartiles. **b** Monthly box charts of MSA at Lampedusa, each box chart displays the median (line inside of each box), the 1st and 3rd quartiles (bottom and top edges of each box), the mean (connected line), and the climatology of F_{DMS} from different estimates inside the same box area.

Lampedusa (see Methods), a central MED site (Fig. S1), from 2005 to 2019. A subset of the available MSA samples was chosen to represent atmospheric marine boundary layer conditions in the MED, hence limiting continental and/or anthropogenic contamination. This was accomplished, according to the back-trajectory analysis (Section 4.4, Fig. S13 and Fig. S14), by selecting samples synchronized only to air masses with a high degree of interaction with seawater when passing through the MED domain.

Figure 6a shows the scatter plot between the observed MSA at Lampedusa and the GPR-F_{DMS} averaged inside the box area comprising grid coordinates 32°–42° N and 05°–20° E (black box in Fig. S1). This black box area has the potential to be the main source region of biogenic emissions arriving at the sampling point during the entire measuring period (2005-2019), revealed by air mass frequency analysis. There is a significant correlation ($r = 0.57$; $p < 0.05$) between GPR-F_{DMS} and MSA atmospheric concentration. The link is mostly driven by seasonality, with low F_{DMS} synced to low MSA concentrations in the winter (Dec-Feb), and a substantial increase in both parameters during summer (Jun–Aug). During the period 2005-2019, the summer average [median] MSA concentration was about 38.3 ± 22.5 [34.3] ng m^{-3}, about 10 orders of magnitude higher than the concentration levels in winter (3.6 ± 2.9 [2.9] ng m^{-3}). Simultaneously to MSA concentration, F_{DMS} increased from 2.1 ± 0.7 [2.0] µmol m^{-2} d^{-1} in winter to 7.0 ± 2.0 [6.7] µmol m^{-2} d^{-1} in summer (nearly three times stronger emissions). Apart from the seasonal minima and maxima, F_{DMS} emissions and MSA concentration are very consistent also during the transition seasons of autumn and spring.

The seasonal changes were examined (Fig. 6b) and compared to those calculated using previously published F_{DMS} climatologies (Rev3 and W20). The seasonal trend of MSA concentrations shows peaks in the summer (Jun–Jul) and declines in the winter (Dec–Feb). This tendency aligns with the seasonal fluctuations in GPR-F_{DMS}, emphasizing the importance of high-resolution datasets in representing the real atmosphere and improving the simulation of DMS-derived aerosols and their associated radiative impacts. Conversely, the Rev3-F_{DMS} peak appears earlier (in May) than the MSA concentration peak, while the W20-F_{DMS} peak appears later (in September), and they are generally less consistent with the MSA seasonal trend (Fig. 6b) compared to GPR.

Discussion

DMS, as the main oceanic source of sulfur aerosol particles in the atmosphere, can affect regional and global climate via altering cloud condensation nuclei concentrations. However, it is not effectively represented in climate models, particularly across regional seas, because precisely predicting the sea-to-air F_{DMS} is challenging. The capacity to capture high-frequency spatial and temporal patterns in DMS and F_{DMS} variations is one of the advantages of generating high-resolution data products. The limited performance of previous methods at fine regional scales is becoming increasingly apparent globally[12,15,17]. This undoubtedly encourages the scientific community to focus on fine-scale regional parameterizations.

In the present work, the GPR model predicts seawater DMS with superior performance than previously published methods, capturing 71% of daily variations on completely different independent test subsets. However, the applied model and the previously published ML methods still struggle to capture high and low DMS concentrations[13,14,17]. Such extreme values are scarce but important considering their potential role in radiative forcing. ML algorithms normally assume uniform distributions, although the majority of oceanic and atmospheric datasets have skewed distributions, with certain values within specific ranges occurring less frequently. As a result, models perform better for frequently represented data points than for rare extremes. In addition, during model training, a group of predictors may dominate because of their strong match with the response, whilst other less-weighted predictors may play a role in shaping extreme values. It will take more research to overcome this challenge to use ML more effectively in oceanic and atmospheric studies.

The major concern that can arise about the proposed dataset regards the limited number of observations available to train the GPR model. A detailed discussion on the suitability of the in-situ DMS concentration dataset for the purpose of training the GPR model can be found in the Supplementary material (Text S1, Figs. S15–S18). Here, we stress out that GPR does not work by interpolating the observations in space or time; instead, it derives a relation between the predicted variable and the considered predictors, using then said relation to re-construct the predicted variable outside the training domain, as a function of the measured predictors. For this reason, GPR cannot be reasonably affected by anomalous observations (*i.e.*, observations not representative of the climatological conditions) that might be non-negligible in a small dataset, as far as the relation between predicted variable and predictors remain consistent. This

condition is evidently respected in the training dataset used here, as shown by the good performance of the model to reconstruct the observations. If anomalous observations had distorted or disrupted the DMS vs. predictors relationship, the model would have failed in evidencing such relationship, resulting in a low or null prediction performance.

The most typical technique to validate a ML model is to keep a separate subset of the data for testing in parallel to the using the normal CV on the training subset[12,20]. However, this method demands enough amount of data to apply the train/test split and exclude the test subset from the model building up. Rather than testing the model only once, many models can be tested iteratively on different portions of the data utilizing the nCV. The nCV allows to train and test a model many times with non-overlapping subsets. This strategy is specifically designed to work with a limited number of data points and to determine whether the model can produce unbiased performance on different test subsets. A potential benefit of the nCV is that it uses all of the available data for model training and testing, which should produce results that are more representative of the data population than using only a portion of the data, as with a train/test split or new dataset. Because the nCV adds significantly more information to the model to learn from it, the method can be used even when there is ample data.

Based on the ML-evidenced relation between DMS concentration and the driving predictors employed, the model has been extended to run over non-sampled areas for constructing the presenting datasets. We underline that for running GPR on a new dataset it is not important to achieve an extensive time or space data coverage in the training dataset, but instead to cover a significant fraction of the DMS and predictors variabilities, which was achieved in this study as shown in Fig. S17. The time frame used in the GPR development is 1999–2012, and the model is used to predict from 1998–2020. The capacity to evaluate the robustness of the model's performance over time is limited by the availability of seawater DMS data, emphasizing the importance of doing more observations in regional waters and overlooked seas. Nevertheless, the correlation between MSA measurements and the predicted F_{DMS} provide confidence that the GPR is picking up the prominent patterns. Interestingly, the F_{DMS}-MSA relationship remained unchanged between the training period of 2005–2012 and the extension period of 2013–2019, when no DMS measurements were accessible (Fig. S18), confirming that the DMS concentration and related fluxes were not biased by the GPR algorithm when operating outside the training period.

It is vital to highlight that the sea-to-air F_{DMS} is predominantly determined by seawater DMS concentration, with DMS gas transfer velocity acting as another regulator (EOF analysis). The choice of the parametrization method of gas transfer velocity impacts the F_{DMS} quantification. We used the Blomquist et al.[46] method, which yields the best match with the MSA (Fig. S19) and is suggested by Bhatti et al.[53] study. However, we report that the use of Goddijn-Murphy et al.[54] and Nightingale et al.[55] parametrizations provided higher F_{DMS} by around 18% and 28%, respectively, compared to Blomquist et al.[46], evaluated over the MED domain on an annual basis (Fig. S20).

The improved dataset outperforms previous offerings in terms of describing mesoscale spatiotemporal variability of DMS concentration and sea-to-air flux; this essential dataset is imperative for long-term studies on marine aerosol and the assessment of its radiative impacts in climate models. Importantly, the GPR model can capture the summer DMS paradox (EOF analysis), which is a characteristic phenomenon of most of the oligotrophic basins. The CHL in the MED sea, as a proxy of phytoplankton activity, do not follow the DMS and MSA concentrations seasonal trend, evidence of the complexity of the MED biogeochemical cycle[31]. This is most likely due to the various mechanisms that contribute to the breakdown of phytoplankton cells and the release of DMS, as well as DMS emission into the atmosphere and oxidation to MSA. When biological activity is at its peak in late winter and early spring due to MED general circulation and deep-water formation and the solar radiation has not yet reached its peak in the summer, DMS release is still modest. When solar radiation reaches its maximum in summer, DMS and F_{DMS} tend to be maximized accordingly. In essence, in warm-oligotrophic environments like the MED sea, the annual cycle appears to be considerably more linked to physical parameters (e.g., solar radiation and surface temperature), whereas biotic variables contribute to quick (high frequency) DMS adjustments. The fact that GPR is able to reconstruct such complex interactions between DMS concentration and its predictors contributes to building confidence on the reliability of the modeled dataset.

Notwithstanding the above pieces of evidence supporting the general reliability of the modeled DMS concentration fields, we cannot rule out some degree of uncertainty due to the limited space-time coverage of DMS observations in the MED basin. For instance, we evidence that no measurements were available for the Adriatic Sea and therefore we invite future users to be careful when extrapolating DMS data for such area, considering its enclosed nature and oceanographic peculiarity.

As an effective tool for predicting DMS, as demonstrated in this study and the previous one in the North Atlantic[12], GPR can be run in different biogeochemical provinces of the global ocean[56], to produce much more reliable products with high spatial resolutions. This necessitates global advances in physical ocean reanalysis as well. Enhanced global oceanic products have the potential to significantly improve the simulation of aerosols originating from DMS even in marine regions with complex morphology and dynamics as well as the resulting regional-scale aerosol-cloud interaction effects.

According to both present and future projections, the MED is a hotspot for global warming because its changes have been more rapid than those of the ocean as a whole[57], having a significant effect and increasing risks on all sectors of the marine environment during the coming decades[58–60]. Reliable representations of marine biogenic emissions, as well as their long-term variations and future scenarios in the context of climate change, are among the key products to reduce aerosol-cloud interaction uncertainty in MED regional climate models.

Methods

Study domain and data sources

Our research domain is the MED Sea, extending from 30° to 46° N and from 06° W to 36.5° E (Fig. S1). Geographically, it is divided into three major subbasins: the Western Basin (WMED), which is connected to the Atlantic Ocean by the Strait of Gibraltar, the Central Basin (CMED), which encompasses the Sicily Channel, the Ionian Sea, and the Adriatic Sea, and the Western basin (WMED), which is connected to the Black Sea by the Dardanelles Strait and the Sea of Marmara. The datasets used for this study are retrieved from satellite products, in-situ data, and model reanalysis in the period 1998–2020. The datasets for this research were obtained from high-resolution satellite products and the Mediterranean Sea Physics Reanalysis. They are briefly described below:

- Daily Level4 (L4) chlorophyll-a concentration (CHL) taken from Copernicus-GlobColour Satellite Observations at 0.042° × 0.042° resolutions.
- Daily photosynthetically available radiation (PAR) data were collected from NASA Ocean Color products such as SeaWiFS (1998–2002), MODIS-Terra (2001–2021), and MODIS-Aqua (2003-2021). SeaWiFS has a spatial resolution of 9 km (0.083° × 0.083°), while MODIS has a resolution of 4 km (0.042° × 0.042°); both are L3 outputs. The data were merged and linearly interpolated before being processed as L4 data.
- The L4 daily SST fields[61,62] were generated by the EU Copernicus Marine Environment Monitoring Service (CMEMS) using satellite estimates, with 0.05° × 0.05° spatial resolution.
- The daily sea surface salinity (SSS) and mixed layer depth (MLD) were taken from the MED Sea physical reanalysis at 1/24° × 1/24° resolution[39]. This data is a multiyear output generated by a numerical system comprised of a hydrodynamic model and a variational data assimilation method developed in the CMEMS framework.

In addition, in-situ observations of surface seawater DMS concentrations were obtained from the Global Surface Seawater DMS Database

(Pacific Marine Environmental Laboratory, PMEL[5]) in the MED sea domain and from Royer et al.[16] campaigns. The sampling points are shown in Fig. S1. All the data including DMS, the aforementioned satellite data and physical reanalysis data were binned into 0.083° × 0.083° (~9 km) daily resolution. The binned dataset was used in the training and nCV validation of the GPR model.

GPR training and the nested cross-validation

We built up the GPR model using the daily binned (0.083° × 0.083°) independent variables (features) as predictors of seawater DMS concentrations. The predictors are CHL, SST, PAR, MLD, SSS, and the day of year (DOY). The use of DOY as a temporal predictor seeks to compensate for the lack of data continuity in observations. GPR[12,20] is an efficient ML algorithm that solves regression problems using a non-parametric kernel-based Bayesian probabilistic strategy[63]. It is powerful on small datasets since contemporary kernel (covariance) functions are readily available[64]. The kernel function used in this study is the exponential that has been defined as the best optimal one in predicting DMS[12] in a previous attempt. In the MED domain, a total of 402 remapped data points, jointly obtained from cruises and fixed stations (Fig. S1). After transforming to the log- scale, of the 402 data points, we eliminated two points that lie below the threshold (mean – three times the standard deviation) and three points that lie above the threshold (mean + three times the standard deviation), which can be considered outliers. Ultimately, 1.2% of the total points have been eliminated and the rest of data points which equal to 397 were used for the model generation.

We evaluated the GPR using the nCV loop to prevent model overfitting and guarantee the unbiased generalization of the trained model (Fig. 1a). In this manner, the dataset is divided into k outer folds approximately of equal size; each outer fold is kept aside for testing, and the remaining k-1 folds are combined and further divided into inner folds for training and cross-validation. We used 5-fold cross-validation in both the inner and outer loops, keeping in mind that the dataset was divided randomly without repetition between the subsets. This means that all the data has been divided into 5 folds (each with about 80 points). Each time in the outer loop we select a fold as a test set, while the remaining four folds are used for training. The four folds are divided once more into five groups to apply the standard cross-validation[12,20] in the inner loop. This procedure will be repeated five times; hence each part of the data will serve as the testing set for once; afterwards, the simulated results when serving as the testing set are aggregated and plotted versus the observed DMS to show the generalization ability of the GPR model. The nCV enables the model to be tested across all independent test subgroups and yield an average performance. The approach is particularly useful in situations when there is a limited amount of data, as in our case, as it enables multiple training and testing of a model utilizing the non-overlapping parts (i.e., folds) of the dataset.

The GPR model was used to generate the long-term (1998–2020) gridded fields of high-resolution (0.083° × 0.083°) daily DMS distributions across the MED sea. Consequently, daily sea-to-air F_{DMS} were calculated using seawater DMS concentrations by GPR and the gas transfer velocity (K_{DMS}), which in turn depends on the surface wind speed (WS) and the DMS diffusivity through seawater (Schmidt number; Sc_{DMS}). The K_{DMS} has been parametrized using the following equation of Blomquist et al.[46]:

$$K_{DMS} = 0.7432 \times (WS)^{1.352} \times (660/Sc_{DMS})^{1/2}$$

The WS is the neutral wind speed at 10 m above the sea surface and has been downloaded from the ECMWF-ERA5 reanalysis dataset[65]. The Sc_{DMS} has been calculated using Saltzman et al.[66] from SST.

MSA sampling and analytical determination

The atmospheric particulate MSA concentration sampling was carried out at the Station for Climate Observations, maintained by ENEA (the National Agency for New Technologies, Energy, and Sustainable Economic Development of Italy) at Lampedusa (35.5 °N, 12.6 °E) in the central Mediterranean (Fig. S1). Particulate matter with aerodynamic diameter lower than 10 μm (PM_{10}) was sampled during 2005-2019 at 24 h time resolution, except the year 2019 which has 48 h time resolution. Detailed description of the applied sampling protocols and analysis methodologies can be found in Becagli et al.[31].

Air mass back-trajectories analysis

We analysed the air mass back-trajectories (BTs) to identify the main areas of the MED domain that can act as a source region of biogenic emissions arrived at the Lampedusa measuring site. The BTs were calculated by running the Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT4) model. The starting height is set to be 100 m AGL, and the backward time is 3 days with an interval of 1 h along each entire trajectory track. The arrival frequency is 3 h (eight tracks per day at 00, 03, 06, 09, 12, 15, 18, 21 UTC) covering the period of MSA measurements (2005–2019). The total number of BTs tracks is 43,824 (5478 days × 8 tracks a day). The frequency of BTs endpoints of all tracks is presented in Fig. S13. This was done by counting the number of endpoints in each 0.25° × 0.25° grid cell and normalized them to the maximum value to find the percentage of endpoints for all grid cells.

To identify the air mass tracks characterized by a high degree of contact with the MED surface water before arriving at Lampedusa, we calculated two indices: the retention ratio of the air mass over the MED seawater (R_0) and the retention ratio of an ocean air mass within the marine boundary layer (R_B). The R_0 values quantify the lifetime of air masses spending over the seawater to the total lifetime whereas, R_B evaluates how such an air mass is confined within the boundary layer height. The boundary layer height datasets at each endpoint were extracted from the hourly ECMWF-ERA5 dataset. The approach and equations are described in detail in Mansour et al.[20]. Ultimately, the marine air masses included in this study are identified if they have $R_0 \geq 0.8$ and $(R_0 + R_B) \geq 1.3$. In this way, we kept the air mass that had a high degree of contact with the MED surface water within the last 3 days before arrival at the Lampedusa site. Fig. S14 displays the frequency of endpoints of the identified marine tracks of the air masses passing across the MED domain. The sea area presenting high marine BTs frequency (more than the 3rd quartile) was identified as the most likely source region of biogenic emissions impacting Lampedusa measured samples during the investigated period (2005–2019). The MSA concentration samples corresponding to marine BTs tracks were selected and compared to DMS flux averaged within this area, delineated by the green box in Fig. S14.

Data availability

The satellite CHL concentrations are available from Copernicus-GlobColour Satellite Observations, accessible at https://doi.org/10.48670/moi-00281. The SST fields were obtained from the Mediterranean Sea - High Resolution L4 Sea Surface Temperature, accessible at https://doi.org/10.48670/moi-00173. The PAR (SeaWiFS, MODIS-Terra, and MODIS-Aqua) were obtained from NASA Ocean Color at https://oceancolor.gsfc.nasa.gov. The SSS and MLD were obtained from Mediterranean Sea Physics Reanalysis, accessible at https://doi.org/10.25423/CMCC/MEDSEA_MULTIYEAR_PHY_006_004_E3R1. The ECMWF ERA5 neutral wind speed data were obtained from https://doi.org/10.24381/cds.adbb2d47. The DMS observations data are available from http://saga.pmel.noaa.gov/dms/ and the Royer et al.[16] cruises are obtained from Martí Galí. The MSA measurements at Lampedusa are available based upon request from the corresponding author.

Code availability

The codes used in this study are available upon request from the corresponding author.

Received: 14 June 2024; Accepted: 30 October 2024;
Published online: 09 November 2024

References

1. Mansour, K. et al. Phytoplankton impact on marine cloud microphysical properties over the Northeast Atlantic Ocean. *J. Geophys. Res.—Atmos.* **127**, e2021JD036355 (2022).
2. Mansour, K. et al. Particulate methanesulfonic acid over the central Mediterranean Sea: Source region identification and relationship with phytoplankton activity. *Atmos. Res.* **237**, 104837 (2020).
3. Mansour, K. et al. Linking marine biological activity to aerosol chemical composition and cloud-relevant properties over the North Atlantic Ocean. *J. Geophys. Res.—Atmos.* **125**, e2019JD032246 (2020).
4. Charlson, R. J., Lovelock, J. E., Andreae, M. O. & Warren, S. G. Oceanic phytoplankton, atmospheric sulfur, cloud albedo and climate. *Nature* **326**, 655–661, https://doi.org/10.1038/326655a0 (1987).
5. Kettle, A. J. et al. 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 (1999).
6. Lana, A. et al. An updated climatology of surface dimethlysulfide concentrations and emission fluxes in the global ocean. *Glob. Biogeochem. Cycles* **25**. https://doi.org/10.1029/2010gb003850 (2011).
7. Hulswar, S. et al. Third revision of the global surface seawater dimethyl sulfide climatology (DMS-Rev3). *Earth Syst. Sci. Data* **14**, 2963–2987 (2022).
8. Simo, R. & Dachs, J. Global ocean emission of dimethylsulfide predicted from biogeophysical data. *Glob. Biogeochem. Cycles* **16**, 26-1-26-10 (2002).
9. Vallina, S. M. & Simo, R. Strong relationship between DMS and the solar radiation dose over the global surface ocean. *Science* **315**, 506–508 (2007).
10. Gali, M., Levasseur, M., Devred, E., Simo, R. & Babin, M. Sea-surface dimethylsulfide (DMS) concentration from satellite data at global and regional scales. *Biogeosciences* **15**, 3497–3519 (2018).
11. Gali, M., Devred, E., Levasseur, M., Royer, S. J. & Babin, M. A remote sensing algorithm for planktonic dimethylsulfoniopropionate (DMSP) and an analysis of global patterns. *Remote Sens. Environ.* **171**, 171–184 (2015).
12. Mansour, K., Decesari, S., Ceburnis, D., Ovadnevaite, J. & Rinaldi, M. Machine learning for prediction of daily sea surface dimethylsulfide concentration and emission flux over the North Atlantic Ocean (1998–2021) *Sci. Total Environ.* **871**, 162123 (2023).
13. Wang, W. L. et al. Global ocean dimethyl sulfide climatology estimated from observations and an artificial neural network. *Biogeosciences* **17**, 5335–5354 (2020).
14. Bell, T. G. et al. Predictability of seawater DMS during the North Atlantic Aerosol and Marine Ecosystem Study (NAAMES). *Front. Marine Sci.* **7** (2021). https://doi.org/10.3389/fmars.2020.596763 (2021).
15. Herr, A. E., Kiene, R. P., Dacey, J. W. H. & Tortell, P. D. Patterns and drivers of dimethylsulfide concentration in the northeast subarctic Pacific across multiple spatial and temporal scales. *Biogeosciences* **16**, 1729–1754, https://doi.org/10.5194/bg-16-1729-2019 (2019).
16. Royer, S. J. et al. A high-resolution time-depth view of dimethylsulphide cycling in the surface sea. *Sci. Rep.* **6**, 32325 (2016).
17. McNabb, B. J. & Tortell, P. D. Improved prediction of dimethyl sulfide (DMS) distributions in the northeast subarctic Pacific using machine-learning algorithms. *Biogeosciences* **19**, 1705–1721 (2022).
18. Tesdal, J. E., Christian, J. R., Monahan, A. H. & von Salzen, K. Evaluation of diverse approaches for estimating sea-surface DMS concentration and air-sea exchange at global scale. *Environ. Chem.* **13**, 390–412 (2016).
19. Bock, J. et al. Evaluation of ocean dimethylsulfide concentration and emission in CMIP6 models. *Biogeosciences* **18**, 3823–3860 (2021).
20. Mansour, K. et al. 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. *Earth Syst. Sci. Data* **16**, 2717–2740 (2024).
21. Pinardi, N. & Masetti, E. Variability of the large scale general circulation of the Mediterranean Sea from observations and modelling: a review. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **158**, 153–174 (2000).
22. Siokou-Frangou, I. et al. Plankton in the open Mediterranean Sea: a review. *Biogeosciences* **7**, 1543–1586 (2010).
23. Basterretxea, G., Font-Munoz, J. S., Salgado-Hernanz, P. M., Arrieta, J. & Hernandez-Carrasco, I. Patterns of chlorophyll interannual variability in Mediterranean biogeographical regions. *Remote Sens. Environ.* **215**, 7–17 (2018).
24. El Hourany, R. et al. Phytoplankton diversity in the Mediterranean sea from satellite data using self-organizing maps. *J. Geophys. Res.: Oceans* **124**, 5827–5843 (2019).
25. Bove, M. C. et al. PM10 source apportionment applying PMF and chemical tracer analysis to ship-borne measurements in the Western Mediterranean. *Atmos. Environ.* **125**, 140–151 (2016).
26. Schembari, C. et al. Source apportionment of PM10 in the Western Mediterranean based on observations from a cruise ship. *Atmos. Environ.* **98**, 510–518 (2014).
27. Mihalopoulos, N., Stephanou, E., Kanakidou, M., Pilitsidis, S. & Bousquet, P. Tropospheric aerosol ionic composition in the Eastern Mediterranean region. *Tellus Ser. B: Chem. Phys. Meteorol.* **49**, 314–326 (1997).
28. Belviso, S. et al. DMS dynamics in the most oligotrophic subtropical zones of the global ocean. *Biogeochemistry* **110**, 215–241 (2012).
29. Polimene, L., Archer, S. D., Butenschön, M. & Allen, J. I. A mechanistic explanation of the Sargasso Sea DMS "summer paradox. *Biogeochemistry* **110**, 243–255 (2012).
30. Galí, M. et al. Spectral irradiance dependence of sunlight effects on plankton dimethylsulfide production. *Limnol. Oceanogr.* **58**, 489–504 (2013).
31. Becagli, S. et al. Relationship between methanesulfonate (MS-) in atmospheric particulate and remotely sensed phytoplankton activity in oligo-mesotrophic central Mediterranean Sea. *Atmos. Environ.* **79**, 681–688 (2013).
32. Besiktepe, S., Tang, K. W., Vila, M. & Simó, R. Dimethylated sulfur compounds in seawater, seston and mesozooplankton in the seas around Turkey. *Deep-Sea Res. I—Oceanogr. Res. Pap.* **51**, 1179–1197 (2004).
33. Vila-Costa, M., Kiene, R. P. & Simó, R. Seasonal variability of the dynamics of dimethylated sulfur compounds in a coastal northwest Mediterranean site. *Limnol. Oceanogr.* **53**, 198–211 (2008).
34. Nguyen, B. C., Belviso, S., Mihalopoulos, N., Gostan, J. & Nival, P. Dimethyl sulfide production during natural phytoplanktonic blooms. *Mar. Chem.* **24**, 133–141 (1988).
35. Speeckaert, G., Borges, A. V., Champenois, W., Royer, C. & Gypens, N. Annual cycle of dimethylsulfoniopropionate (DMSP) and dimethylsulfoxide (DMSO) related to phytoplankton succession in the Southern North Sea. *Sci. Total Environ.* **622**, 362–372 (2018).
36. Simó, R. et al. The quantitative role of microzooplankton grazing in dimethylsulfide (DMS) production in the NW Mediterranean. *Biogeochemistry* **141**, 125–142 (2018).
37. Belviso, S. et al. Production of dimethylsulfonium propionate (DMSP) and dimethylsulfide (DMS) by a microbial food web. *Limnol. Oceanogr.* **35**, 1810–1821 (1990).
38. Stefels, J. Physiological aspects of the production and conversion of DMSP in marine algae and higher plants. *J. Sea Res.* **43**, 183–197 (2000).
39. Escudier, R. et al. A high resolution reanalysis for the Mediterranean sea. *Front. Earth Sci.* **9**. https://doi.org/10.3389/feart.2021.702285 (2021).

40. Hannachi, A., Jolliffe, I. T. & Stephenson, D. B. Empirical orthogonal functions and related techniques in atmospheric science: A review. *Int. J. Climatol.* **27**, 1119–1152 (2007).
41. Guo, W., Sun, Q., Wang, S. & Zhang, Z. Characterizing spatio-temporal variations of dimethyl sulfide in the Yellow and East China Sea based on BP neural network. *Front. Marine Sci.* **11** https://doi.org/10.3389/fmars.2024.1394502 (2024).
42. Zhuang, G. C., Yang, G. P., Yu, J. A. & Gao, Y. A. Production of DMS and DMSP in different physiological stages and salinity conditions in two marine algae. *Chin. J. Oceanol. Limnol.* **29**, 369–377 (2011).
43. Speeckaert, G., Borges, A. V. & Gypens, N. Salinity and growth effects on dimethylsulfoniopropionate (DMSP) and dimethylsulfoxide (DMSO) cell quotas of Skeletonema costatum, Phaeocystis globosa and Heterocapsa triquetra. *Estuarine Coastal Shelf Sci.* **226**, 106275 (2019).
44. Salgado, P., Kiene, R., Wiebe, W. & Magalhaes, C. Salinity as a Regulator of DMSP Degradation in Ruegeria pomeroyi DSS-3. *J. Microbiol.* **52**, 948–954 (2014).
45. Thariath, D. V., Divakaran, D. & Chenicherry, S. Influence of salinity on the dimethylsulphoniopropionate production from Prymnesium simplex. *Sustain. Environ. Res.* **29**, 17 (2019).
46. Blomquist, B. W. et al. Wind Speed and Sea State Dependencies of Air-Sea Gas Transfer: Results From the High Wind Speed Gas Exchange Study (HiWinGS). *J. Geophys. Res.-Oceans* **122**, 8034–8062 (2017).
47. Belviso, S., Sciandra, A. & Copin-Montegut, C. Mesoscale features of surface water DMSP and DMS concentrations in the Atlantic Ocean off Morocco and in the Mediterranean Sea. *Deep-Sea Res. I—Oceanogr. Res. Pap.* **50**, 543–555 (2003).
48. Millot, C. & Taupier-Letage, I. *The Mediterranean Sea* (ed. Saliot, A.) 29–66 (Springer, 2005).
49. Messié, M. & Chavez, F. Global modes of sea surface temperature variability in relation to regional climate indices. *J. Clim.* **24**, 4314–4331 (2011).
50. Jiang, Q. F., Smith, R. B. & Doyle, J. The nature of the mistral: Observations and modelling of two MAP events. *Q. J. R. Meteorol. Soc.* **129**, 857–875 (2003).
51. Renault, L. et al. Coupled atmosphere-ocean-wave simulations of a storm event over the Gulf of Lion and Balearic Sea. *J. Geophys. Res.—Oceans* **117**. https://doi.org/10.1029/2012jc007924 (2012).
52. Ziv, B., Saaroni, H. & Alpert, P. The factors governing the summer regime of the eastern Mediterranean. *Int. J. Climatol.* **24**, 1859–1871 (2004).
53. Bhatti, Y. et al. The sensitivity of Southern Ocean atmospheric dimethyl sulfide (DMS) to modeled oceanic DMS concentrations and emissions. *Atmos. Chem. Phys.* **23**, 15181–15196 (2023).
54. Goddijn-Murphy, L., Woolf, D. K. & Marandino, C. Space-based retrievals of air-sea gas transfer velocities using altimeters: Calibration for dimethyl sulfide. *J. Geophys. Res.—Oceans* **117**. https://doi.org/10.1029/2011jc007535 (2012).
55. Nightingale, P. D. et al. In situ evaluation of air-sea gas exchange parameterizations using novel conservative and volatile tracers. *Glob. Biogeochem. Cycles* **14**, 373–387 (2000).
56. Reygondeau, G. et al. Dynamic biogeochemical provinces in the global ocean. *Glob. Biogeochem. Cycles* **27**, 1046–1058 (2013).
57. Cos, J. et al. The Mediterranean climate change hotspot in the CMIP5 and CMIP6 projections. *Earth Syst. Dyn.* **13**, 321–340 (2022).
58. Kim, G. U., Seo, K. H. & Chen, D. L. Climate change over the Mediterranean and current destruction of marine ecosystem. *Sci. Rep.* **9**, 18813 (2019).
59. Cramer, W. et al. Climate change and interconnected risks to sustainable development in the Mediterranean. *Nat. Clim. Change* **8**, 972–980 (2018).
60. Liquete, C., Piroddi, C., Macías, D., Druon, J. N. & Zulian, G. Ecosystem services sustainability in the Mediterranean Sea: assessment of status and trends using multiple modelling approaches. *Sci. Rep.* **6**, 34162 (2016).
61. Pisano, A., Nardelli, B. B., Tronconi, C. & Santoleri, R. The new Mediterranean optimally interpolated pathfinder AVHRR SST Dataset (1982-2012). *Remote Sens. Environ.* **176**, 107–116 (2016).
62. Merchant, C. J. et al. Satellite-based time-series of sea-surface temperature since 1981 for climate applications. *Sci. Data* **6**, 223 (2019).
63. Williams, C. K. I. & Rasmussen, C. E. Gaussian processes for regression. *Adv. Neural Inf. Process. Syst. 8: Proc. 1995 Conf.* **8**, 514–520 (1996).
64. Verrelst, J. et al. Spectral band selection for vegetation properties retrieval using Gaussian processes regression. *Int. J. Appl. Earth Observation Geoinf.* **52**, 554–567 (2016).
65. Hersbach, H. et al. The ERA5 global reanalysis. *Q. J. R. Meteorological Soc.* **146**, 1999–2049 (2020).
66. Saltzman, E. S., King, D. B., Holmen, K. & Leck, C. Experimental determination of the diffusion coefficient of dimethylsulfide in water. *J. Geophys. Res.-Oceans* **98**, 16481–16486 (1993).

Acknowledgements

Funding for this study was provided by the European Commission's EU Horizon 2020 Framework program, project FORCeS (grant no. 821205), and the European Union's Horizon, project CleanCloud (Grant No. 101137639). We gratefully acknowledge the EU Copernicus Marine Environment Monitoring Service (CMEMS) for the provision of satellite data and Mediterranean physics reanalysis, the Copernicus climate change service (C3S) for providing ERA5 reanalysis data, the PMEL project for the observed DMS measurements and the NOAA Air Resources Laboratory (ARL) for the provision of the HYSPLIT transport and dispersion model. Measurements at Lampedusa were partly supported by the Italian Ministry for University and Research through the NextData and Ritmare projects. We thank Damiano Sferlazzo and Francesco Monteleone for the scientific and technical support of the aerosol sampling at Lampedusa. We thank Martí Galí for providing DMS observations and his insightful comments on the results of the manuscript.

Author contributions

K.M. conceptualized and designed the study. S.B. provided the MSA data at Lampedusa. K.M. organized the datasets, constructed the models, analysed the data, and visualized the results. K.M. wrote the manuscript under the supervision of M.R. All authors contributed to the results investigation, manuscript revision, reading and editing and have approved the final version of the manuscript.

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.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41612-024-00830-y.

Correspondence and requests for materials should be addressed to Karam Mansour.