Supplementary Material

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

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

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

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

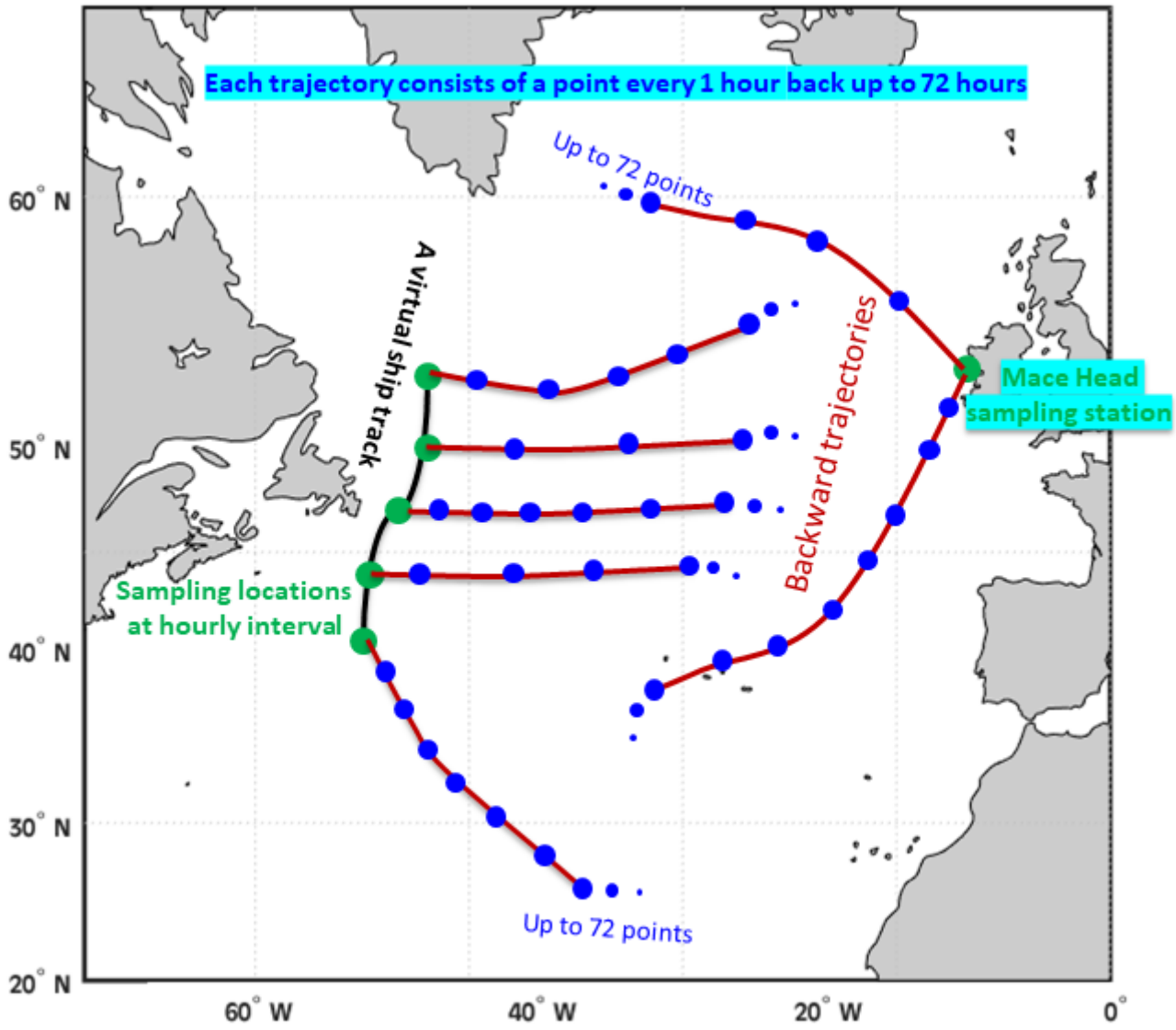

Figure S1: Schematic diagram of the air mass back-trajectories calculated at Mace Head and for NAAMES cruises (a virtual cruise track is represented by the black line). Each trajectory consists of 73 points (green as a start and blue points are backward hours). The predictors have been extracted along each track to consider the air mass history.

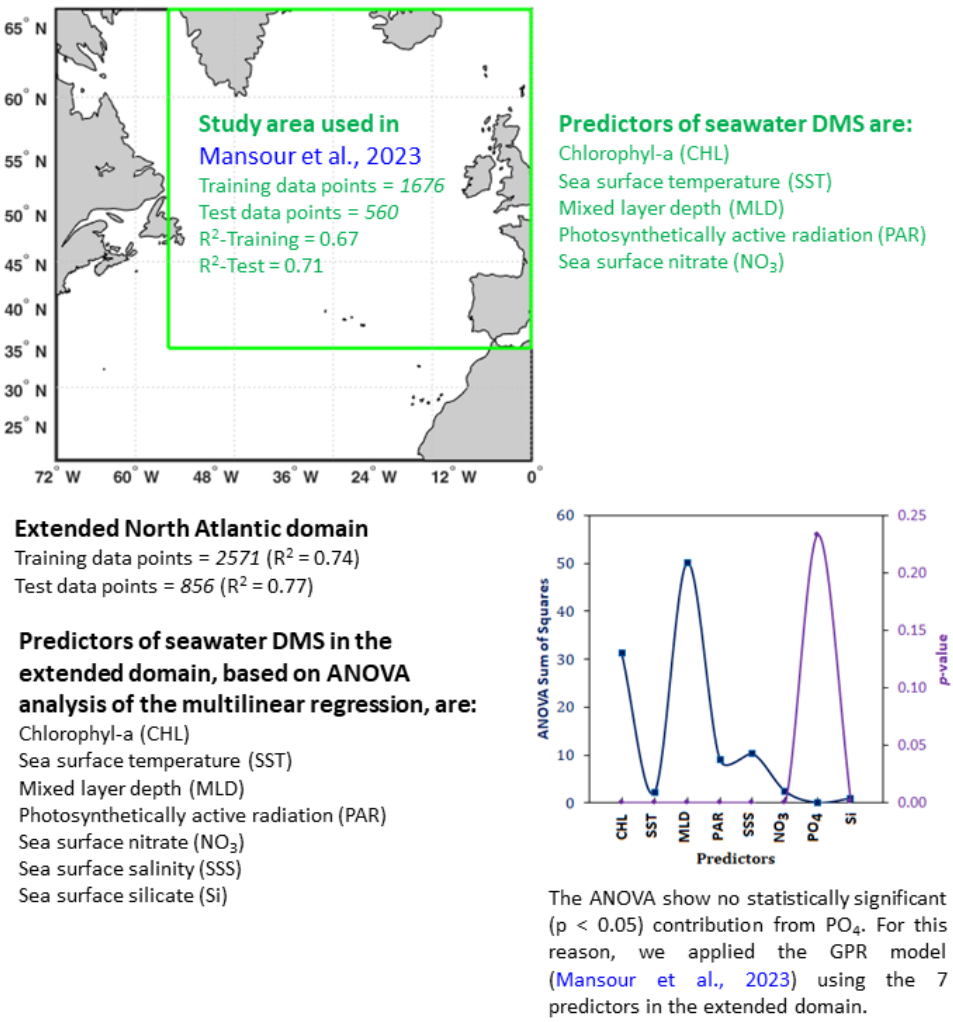

Figure S2: The main differences between the North Atlantic domain used in the present study and the domain used in Mansour et al. 2023 (green box areas).

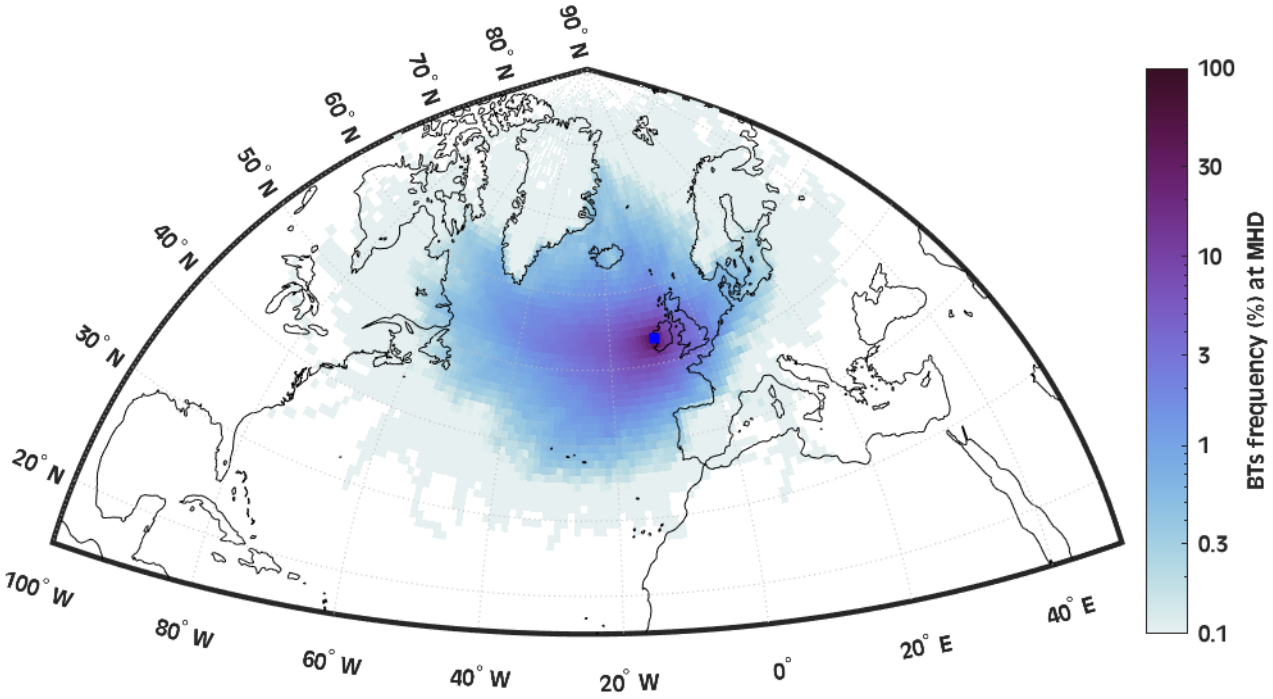

Figure S3: Spatial distributions of the BTs endpoints arriving at Mace Head from Jan 2009 to Jun 2018, displaying where the majority of endpoints are located. The total number of trajectory endpoints was counted in each 1°×1° grid cell and normalized to the maximum value as a percentage.

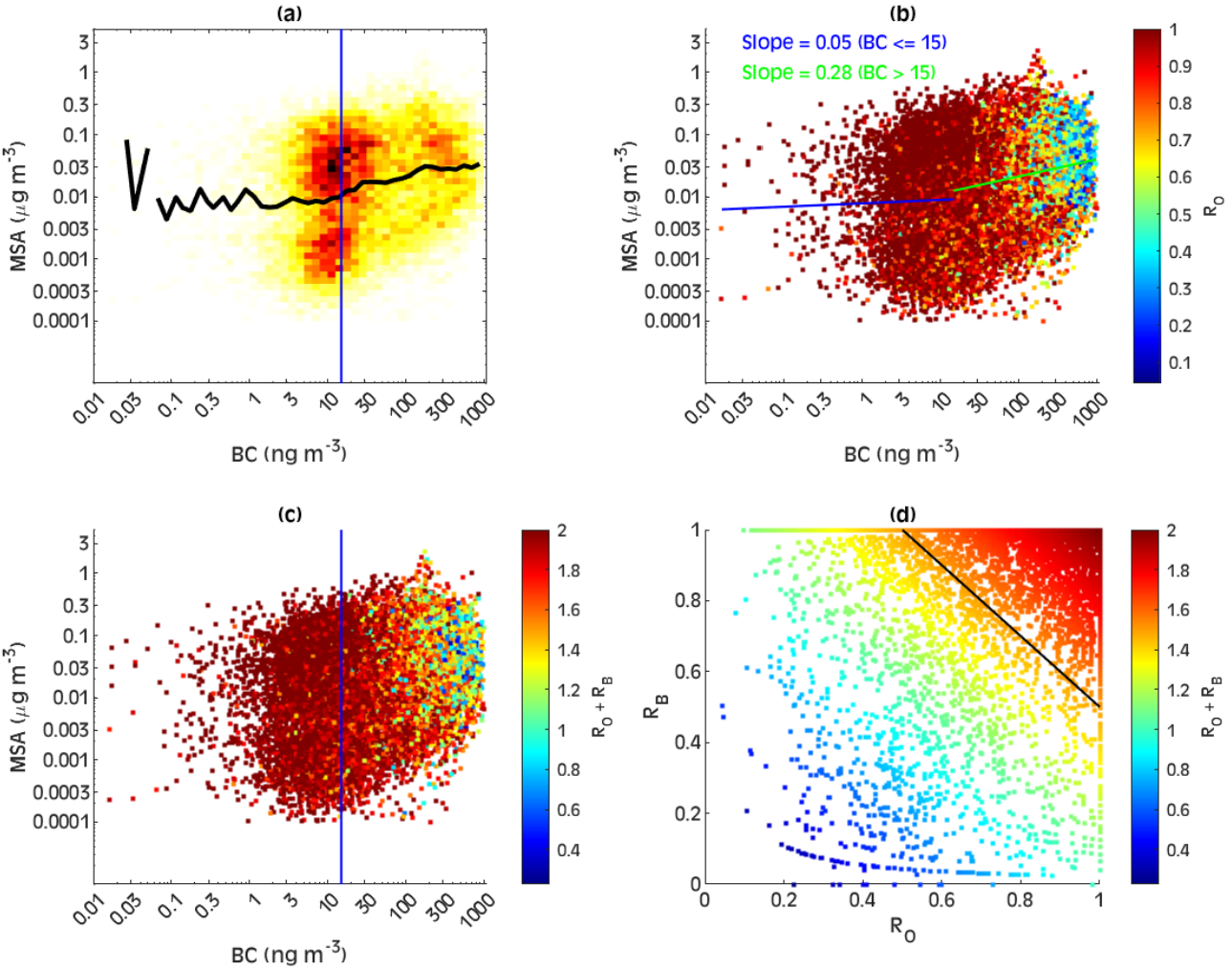

Figure S4: (a) Joint probability histograms of BC-MSA at MHD where darker colors indicate higher probability; the thick black line shows the mean MSA at each BC bin, illustrating MSA behavior as BC changes. The vertical blue line represents the BC value of 15 ng m^{-3}. (b) Scatter plot between MSA and BC where the color scale represents the R_O values. (c) Scatter plot between MSA and BC where the color scale represents the R_O+ R_B values. (d) Scatter plot between R_O and R_O where the color scale represents the sum of them; the points above the diagonal black line have been selected as representative of marine conditions.

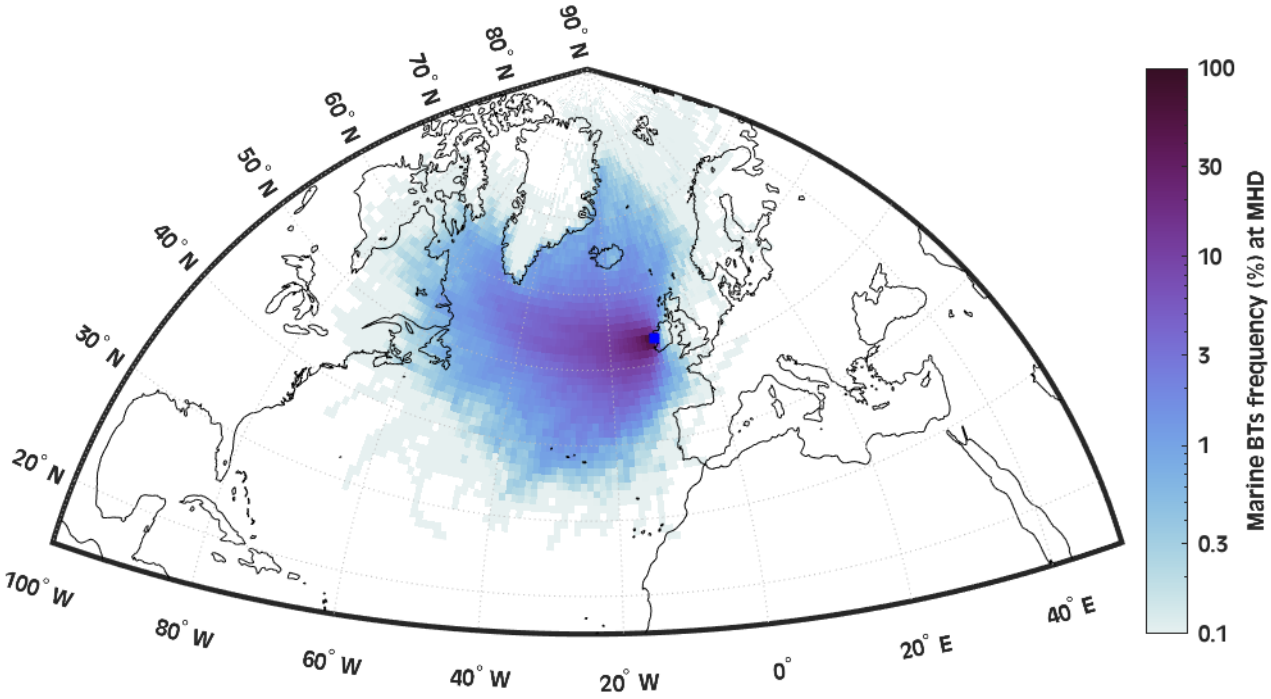

Figure S5: Spatial distributions of the selected marine BTs endpoints arriving at Mace Head from Jan 2009 to Jun 2018, displaying where the majority of endpoints are located. The total number of trajectory endpoints was counted in each 1°×1° grid cell and normalized to the maximum value as a percentage.

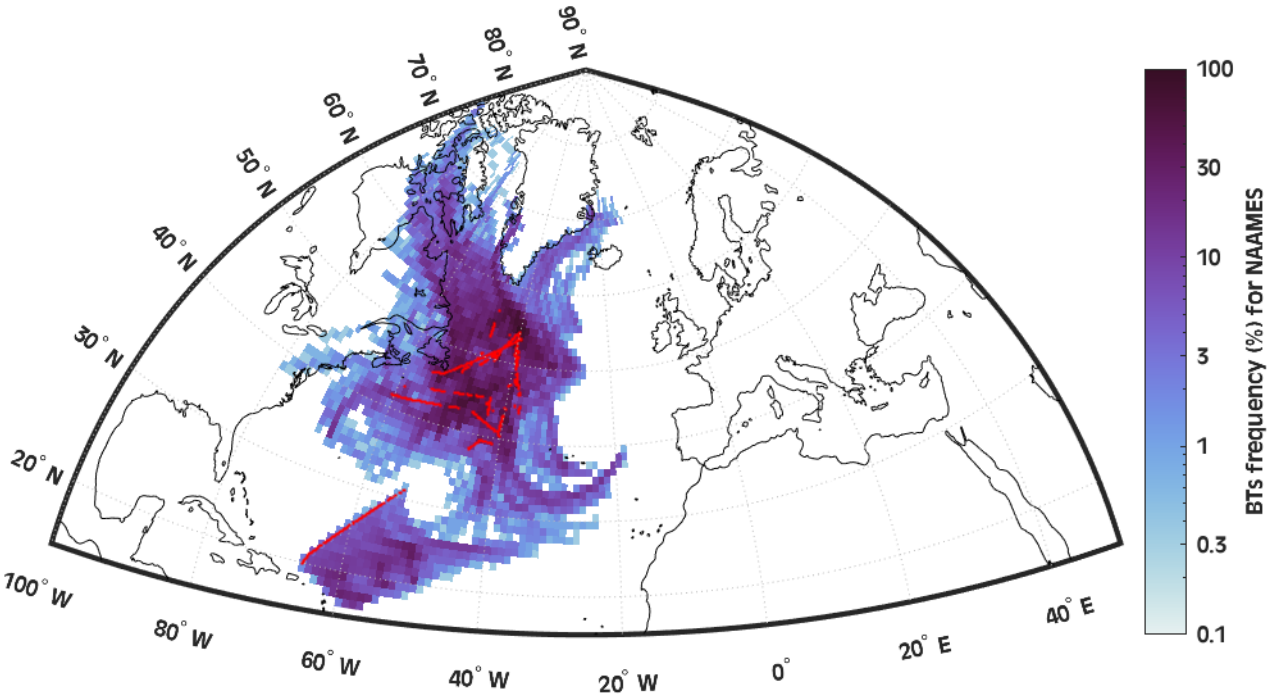

Figure S6: Spatial distributions of the BTs endpoints arriving at NAAMES cruises, displaying where the majority of endpoints are located. The total number of trajectory endpoints was counted in each 1°×1° grid cell and normalized to the maximum value as a percentage.

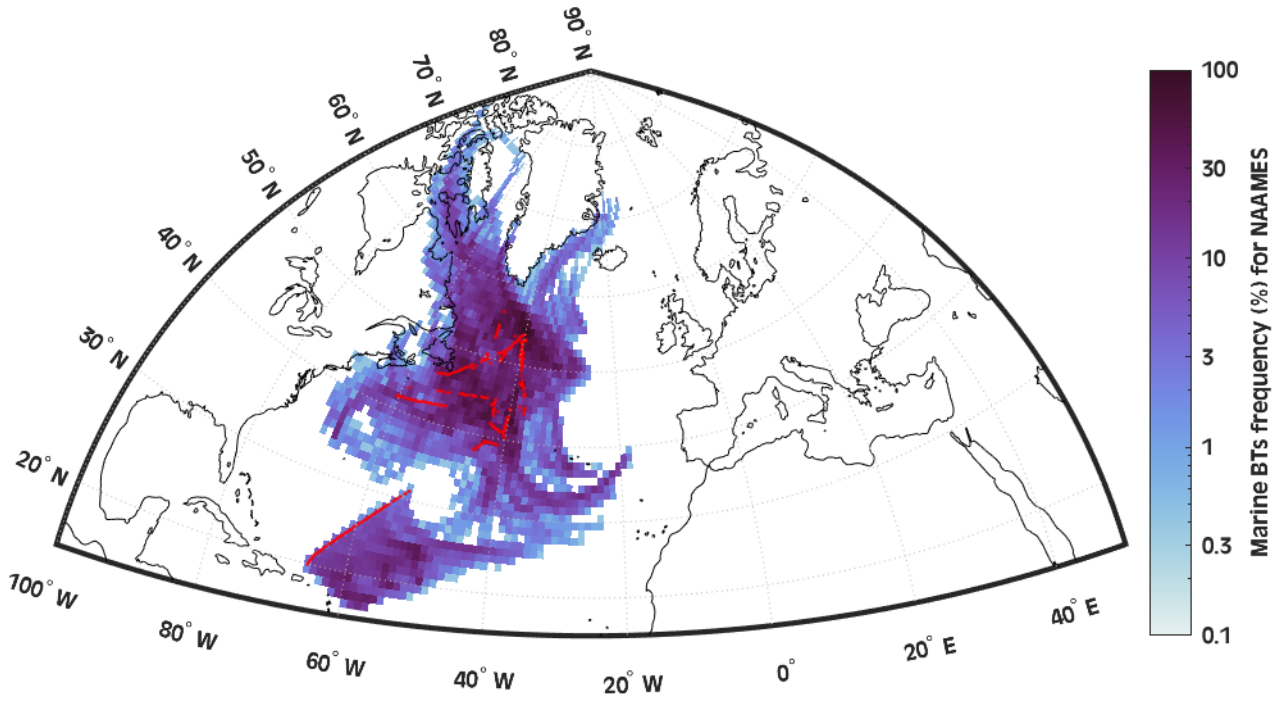

Figure S7: Spatial distributions of the marine BTs endpoints arriving at NAAMES cruises, displaying where the majority of endpoints are located. The total number of trajectory endpoints was counted in each 1°×1° grid cell and normalized to the maximum value as a percentage.

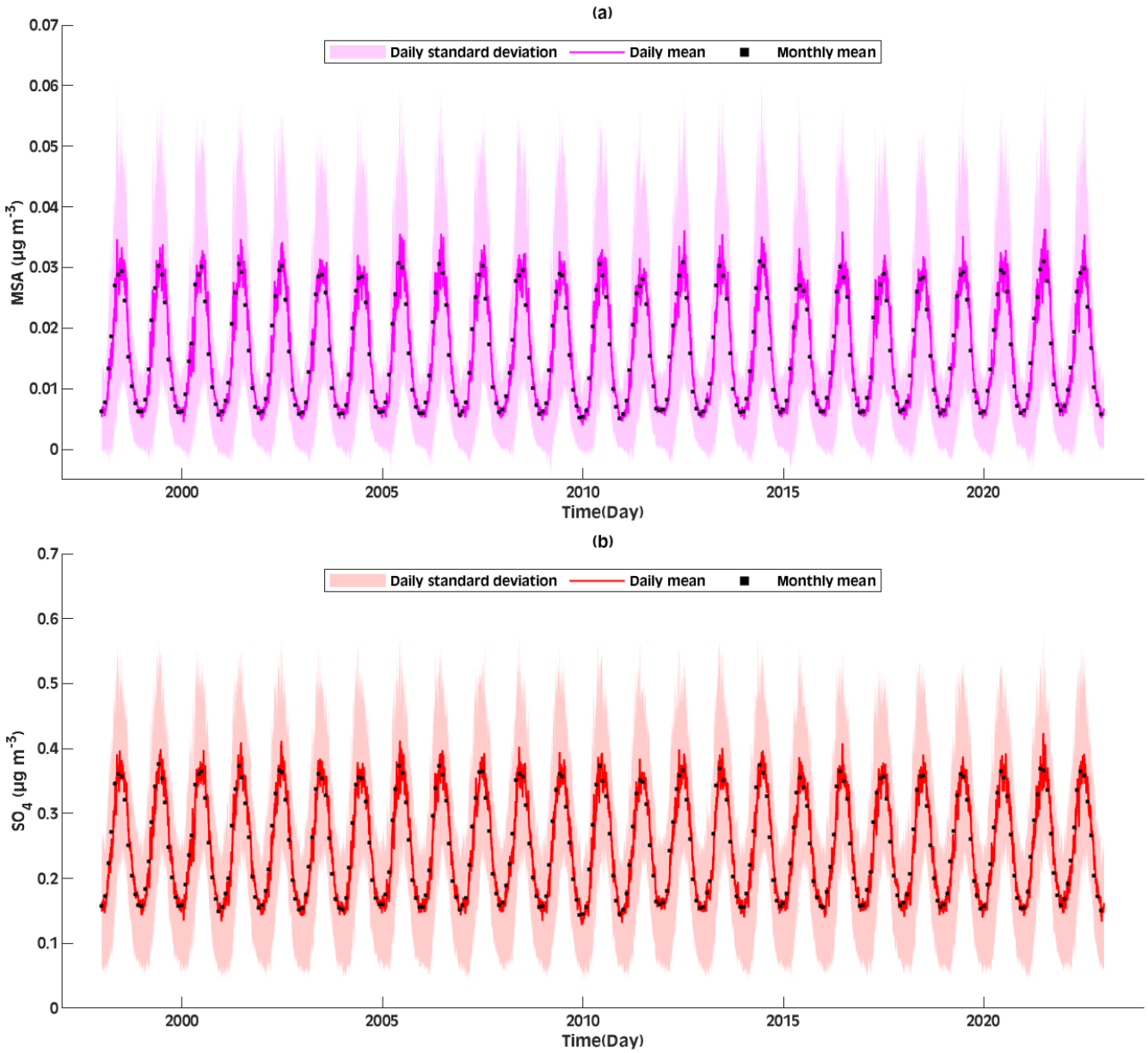

Figure S8: Daily time series of (a) MSA and (b) SO_4 over the entire NA domain obtained by GPR in 1998–2022. The shaded area displays ±1 spatial standard deviation and the black dots represent the monthly mean.

Model Type	Preset	Hyperparameters if any	Cross-validation		Test	
			RMSE	R^2	RMSE	R^2
Multilinear	Linear		0.408	0.74	0.392	0.76
Support Vector Machines	Linear		0.409	0.74	0.393	0.76
	Quadratic		0.398	0.75	0.384	0.77
	Cubic		0.390	0.76	0.376	0.78
	Fine Gaussian	Kernel scale = 0.61	0.410	0.74	0.399	0.75
	Medium Gaussian	Kernel scale = 2.4	0.382	0.77	0.370	0.79
	Coarse Gaussian	Kernel scale = 9.8	0.395	0.76	0.380	0.78
Regression Ensemble	Boosted	Minimum leaf size = 8 Number of learners = 30	0.399	0.75	0.386	0.77
	Bagged	Minimum leaf size = 8 Number of learners = 30	0.374	0.78	0.360	0.80
Gaussian Process Regression	Squared Exponential		0.383	0.77	0.370	0.79
	Matern 5/2		0.377	0.78	0.364	0.79
	Exponential		0.366	0.79	0.350	0.81
	Rational Quadratic		0.362	0.79	0.347	0.81
Neural Networks	Narrow	Number of fully connected layers = 1 First layer size = 10	0.388	0.76	0.374	0.78
	Medium	Number of fully connected layers = 1 First layer size = 25	0.387	0.77	0.379	0.78
	Wide	Number of fully connected layers = 1 1st layer size = 100	0.410	0.74	0.390	0.76
	Bi-layered	Number of fully connected layers = 2 1st layer size = 10 2nd layer size = 10	0.389	0.76	0.378	0.78
	Tri-layered	Number of fully connected layers = 3 1st layer size = 10 2nd layer size = 10 3rd layer size = 10	0.386	0.77	0.373	0.78

Table S1: Evaluation metrics for cross/validation and test datasets of machine learning models and the multilinear model trained to estimate MSA concentrations. Shaded cells represent the best performance from each type.

Model Type	Preset	Hyperparameters if any	Cross-validation		Test	
			RMSE	R^2	RMSE	R^2
Multilinear	Linear		0.324	0.53	0.318	0.55
Support Vector Machines	Linear		0.325	0.53	0.319	0.54
	Quadratic		0.318	0.55	0.310	0.57
	Cubic		0.311	0.57	0.304	0.59
	Fine Gaussian	Kernel scale = 0.61	0.299	0.60	0.293	0.61
	Medium Gaussian	Kernel scale = 2.4	0.305	0.58	0.296	0.61
	Coarse Gaussian	Kernel scale = 9.8	0.315	0.56	0.310	0.57
Regression Ensemble	Boosted	Minimum leaf size = 8 Number of learners = 30	0.309	0.57	0.303	0.59
	Bagged	Minimum leaf size = 8 Number of learners = 30	0.297	0.60	0.283	0.64
Gaussian Process Regression	Squared Exponential		0.305	0.58	0.299	0.60
	Matern 5/2		0.303	0.59	0.295	0.61
	Exponential		0.290	0.62	0.280	0.65
	Rational Quadratic		0.282	0.64	0.272	0.67
Neural Networks	Narrow	Number of fully connected layers = 1 First layer size = 10	0.311	0.57	0.301	0.59
	Medium	Number of fully connected layers = 1 First layer size = 25	0.311	0.57	0.300	0.60
	Wide	Number of fully connected layers = 1 1st layer size = 100	0.322	0.53	0.304	0.59
	Bi-layered	Number of fully connected layers = 2 1st layer size = 10 2nd layer size = 10	0.313	0.56	0.301	0.59
	Tri-layered	Number of fully connected layers = 3 1st layer size = 10 2nd layer size = 10 3rd layer size = 10	0.314	0.56	0.307	0.58

Table S2: Same as Table S1, but for SO_4 concentrations.