Supplement of Earth Syst. Sci. Data, 16, 2717–2740, 2024
https://doi.org/10.5194/essd-16-2717-2024-supplement

Open Access Earth System Science Data

Supplement of

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 et al.

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

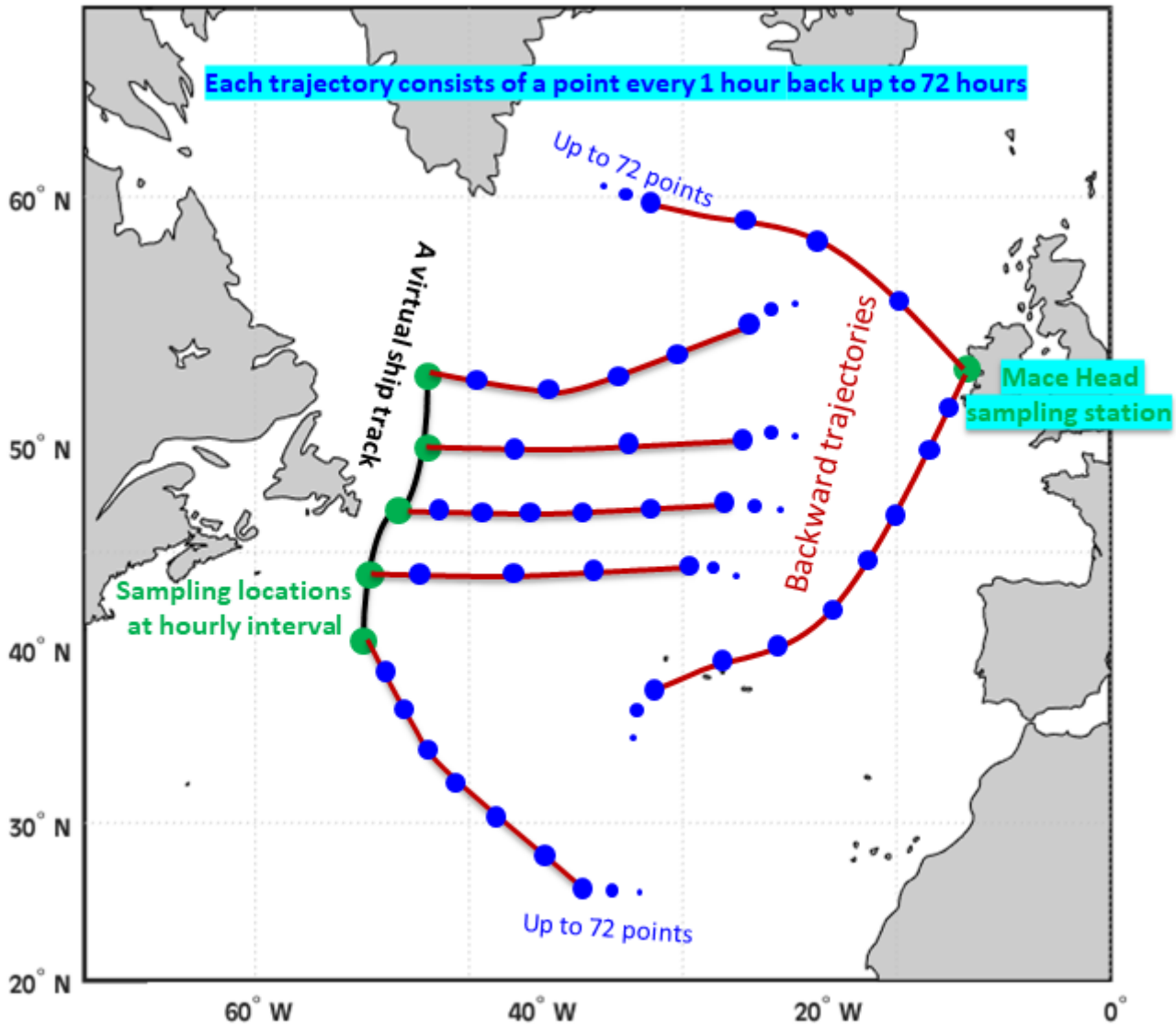

Figure S1: Schematic diagram of the air mass back-trajectories calculated at Mace Head and for NAAMES cruises (a virtual cruise track is represented by the black line). Each trajectory consists of 73 points (green as a start and blue points are backward hours). The predictors have been extracted along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

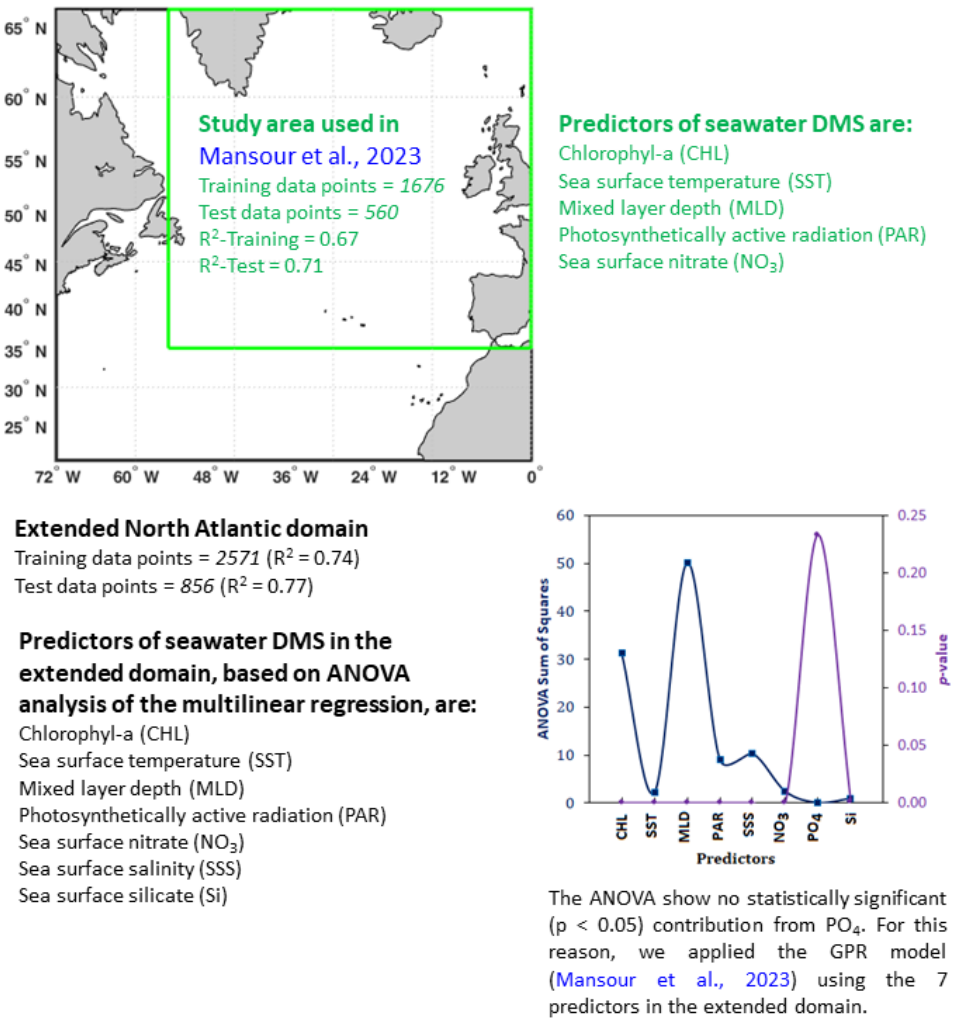

Figure S2: The main differences between the North Atlantic domain used in the present study and the domain used in Mansour et al. 2023 (green box areas).

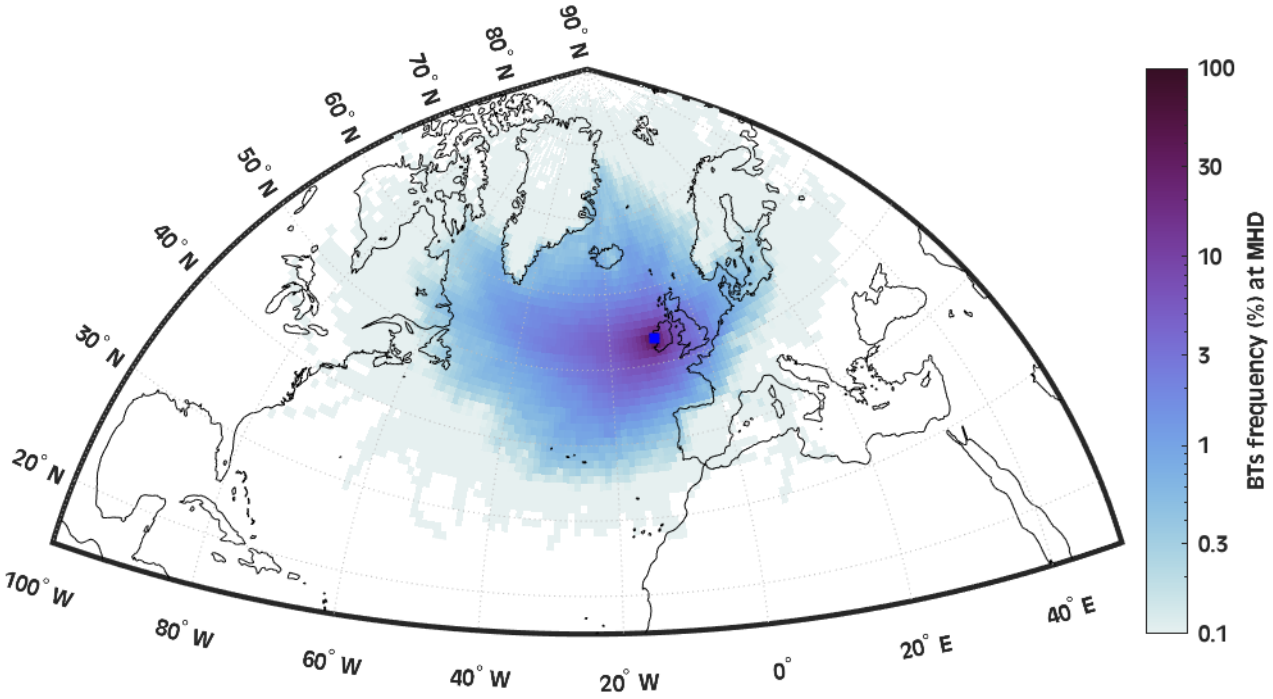

Figure S3: Spatial distributions of the BTs endpoints arriving at Mace Head from Jan 2009 to Jun 2018, displaying where the majority of endpoints are located. The total number of trajectory endpoints was counted in each 1°×1° grid cell and normalized to the maximum value as a percentage.

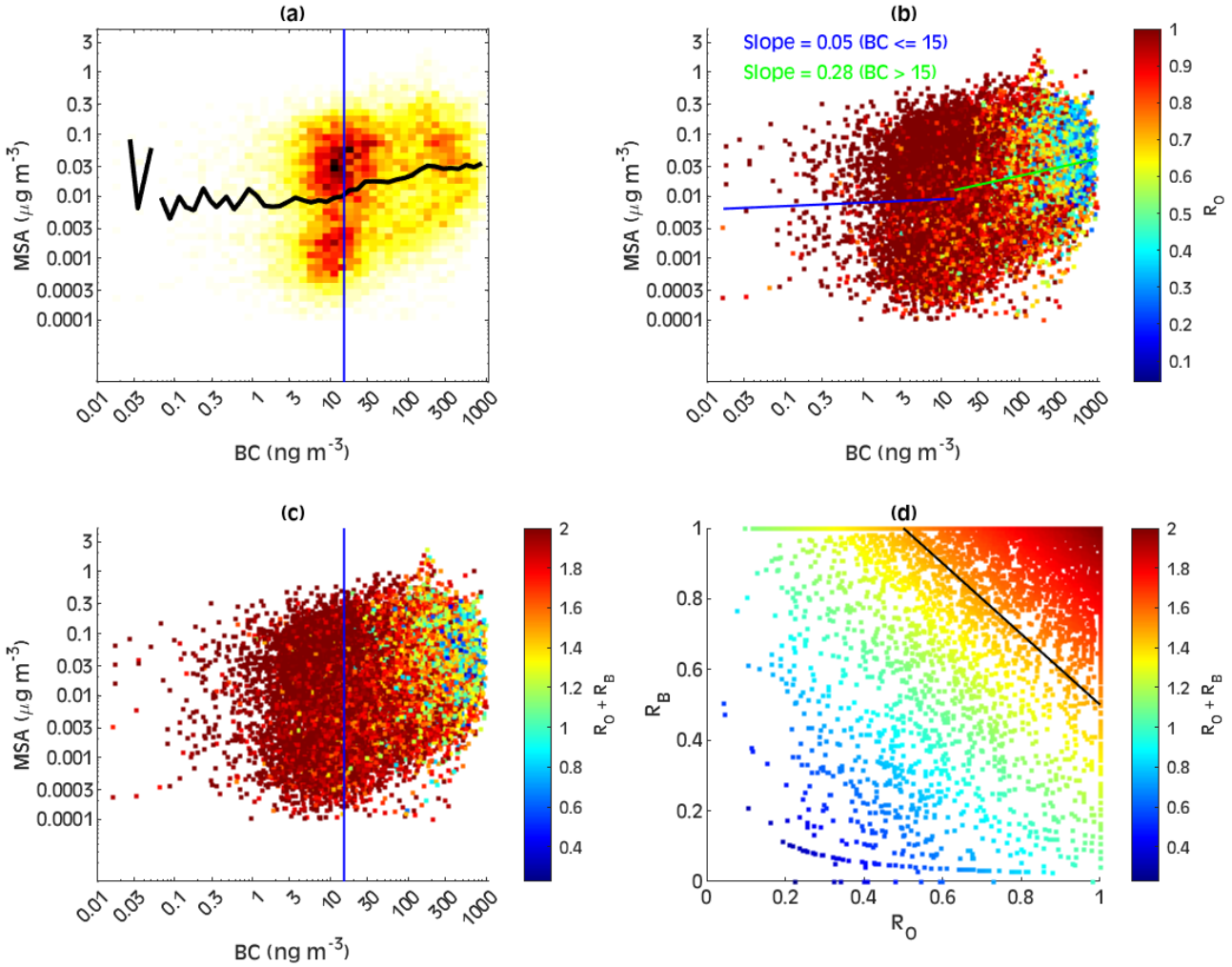

Figure S4: (a) Joint probability histograms of BC-MSA at MHD where darker colors indicate higher probability; the thick black line shows the mean MSA at each BC bin, illustrating MSA behavior as BC changes. The vertical blue line represents the BC value of 15 ng m^{-3}. (b) Scatter plot between MSA and 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.

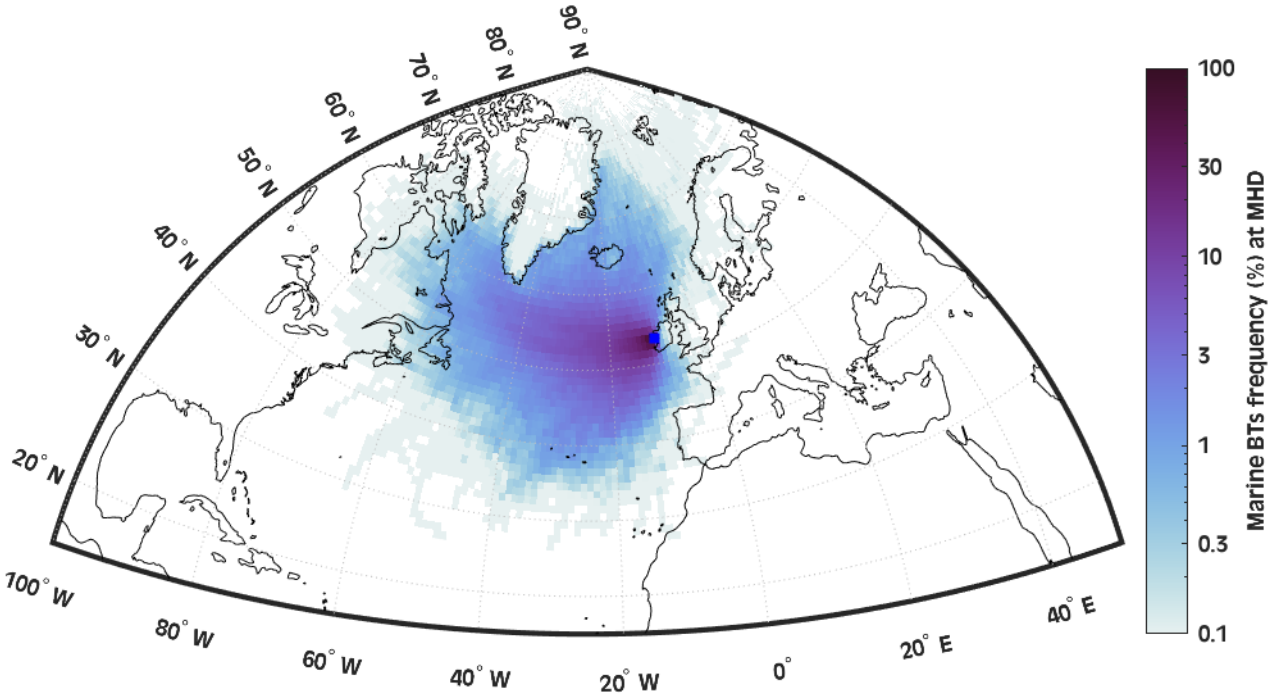

Figure S5: Spatial distributions of the selected marine BTs endpoints arriving at Mace Head from Jan 2009 to Jun 2018, displaying where the majority of endpoints are located. The total number of trajectory endpoints was counted in each 1°×1° grid cell and normalized to the maximum value as a percentage.

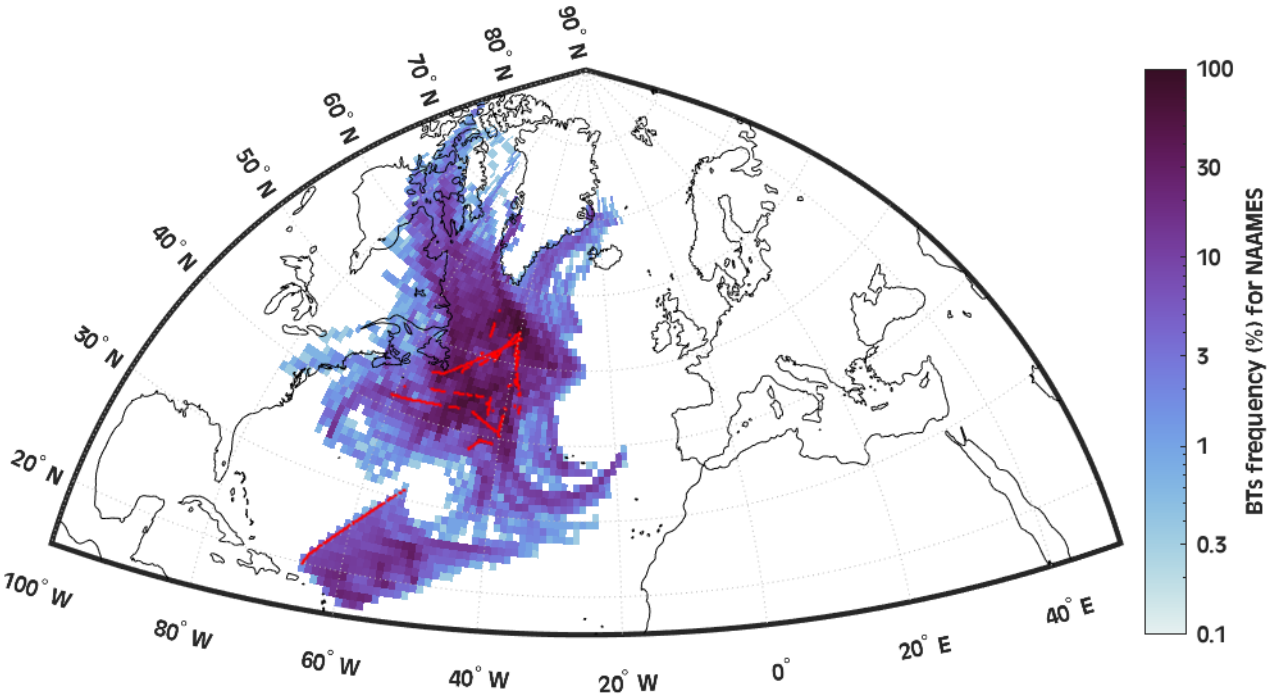

Figure S6: Spatial distributions of the BTs endpoints arriving at NAAMES cruises, displaying where the majority of endpoints are located. The total number of trajectory endpoints was counted in each 1°×1° grid cell and normalized to the maximum value as a percentage.

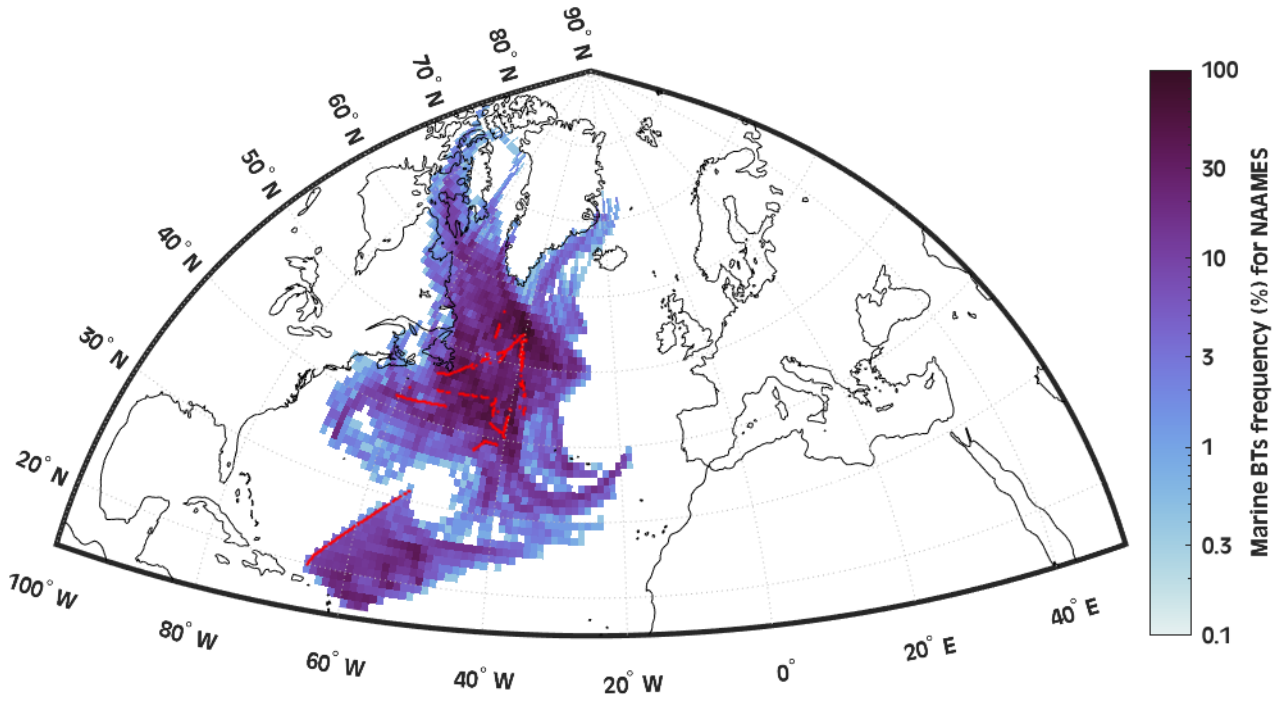

Figure S7: Spatial distributions of the marine BTs endpoints arriving at NAAMES cruises, displaying where the majority of endpoints are located. The total number of trajectory endpoints was counted in each 1°×1° grid cell and normalized to the maximum value as a percentage.

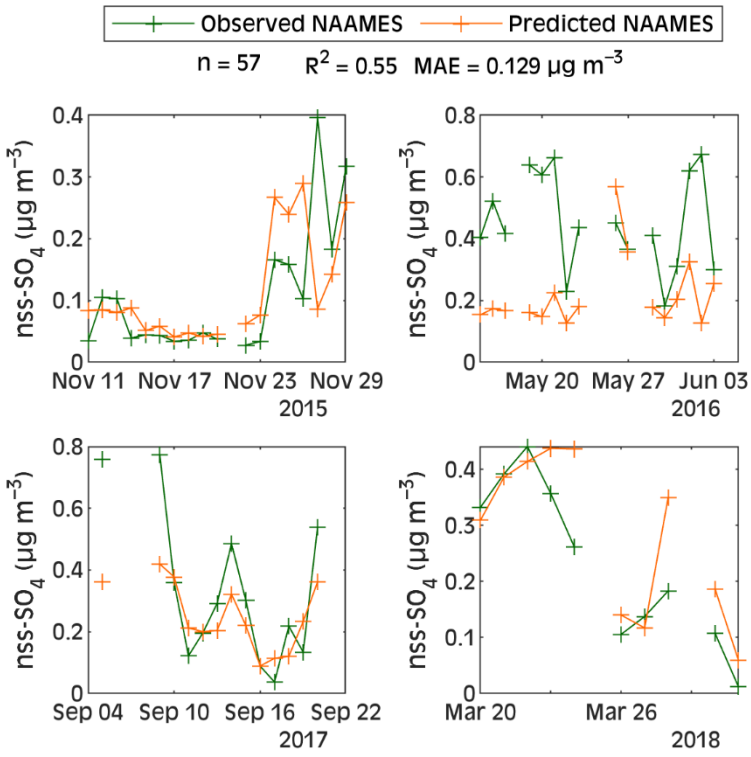

Figure S8: Comparison between daily observed and GPR-predicted nss-$SO_4^=$ during the four NAAMES campaigns. The GPR was trained on the MHD data and tested on the NAAMES data. R^2 is computed in a logarithmic space, whereas MAE is computed on a normal scale.

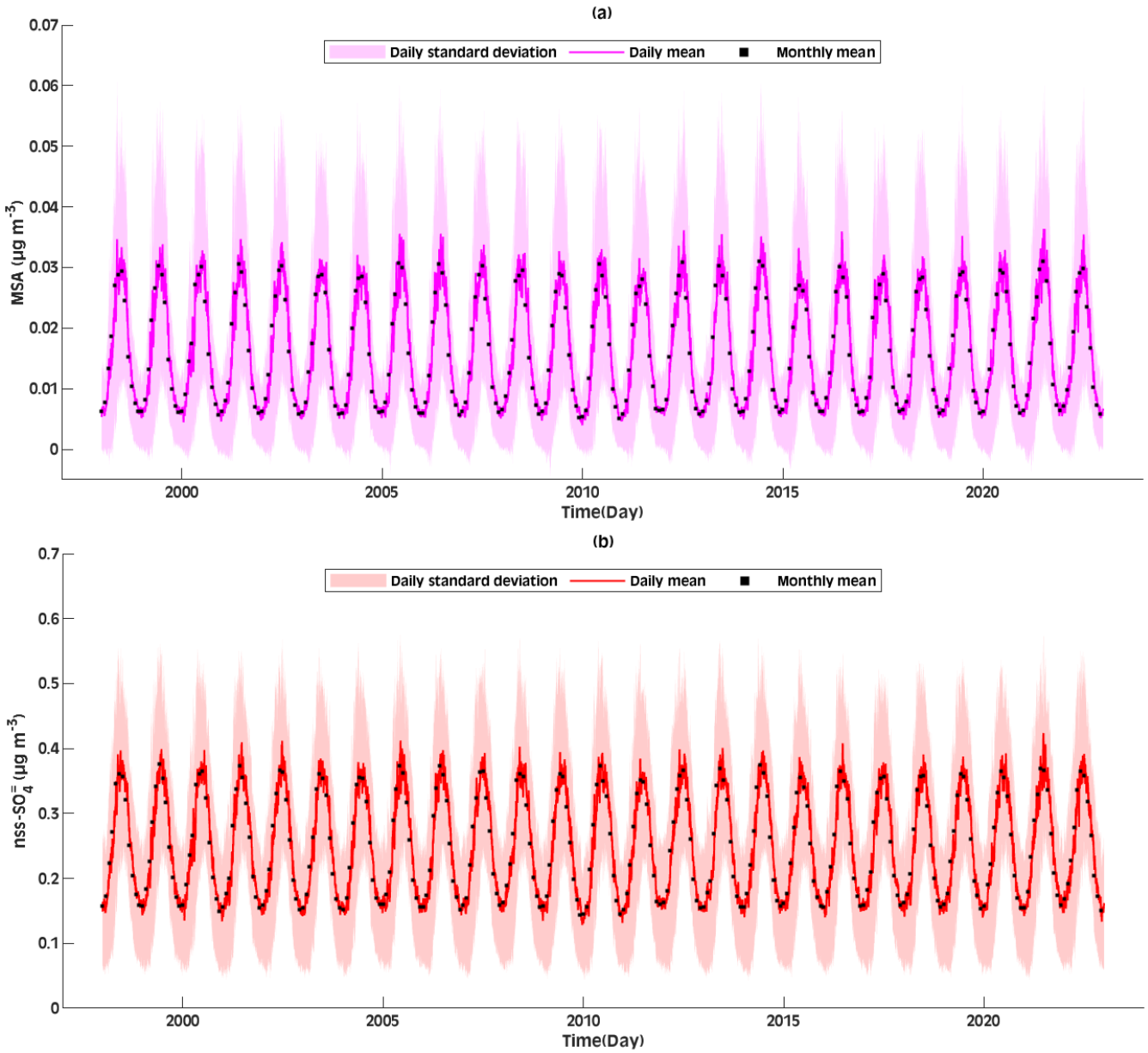

Figure S9: Daily time series of (a) MSA and (b) nss-$SO_4^=$ over the entire NA domain obtained by GPR in 1998–2022. The shaded area displays ±1 spatial standard deviation and the black dots represent the monthly mean.

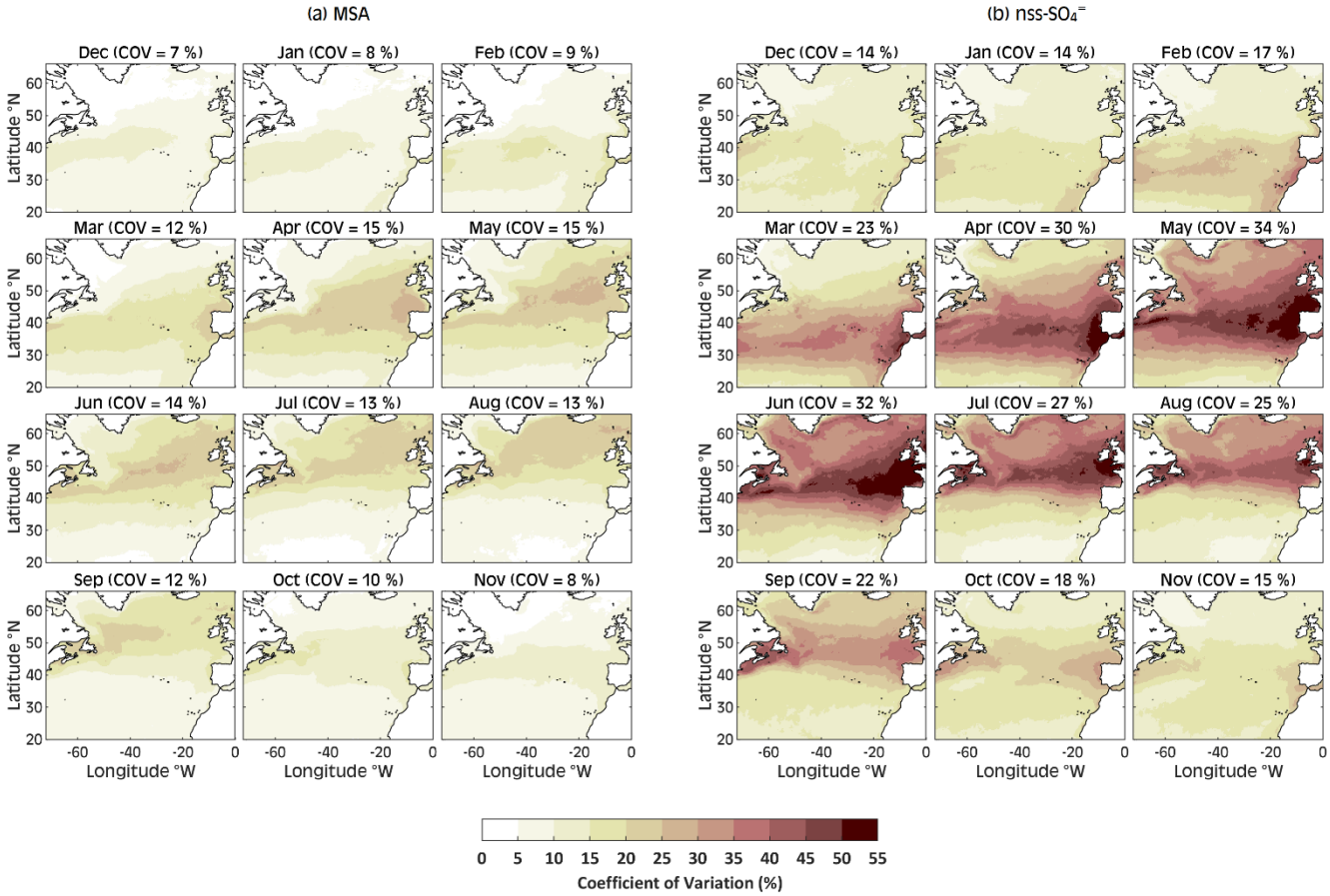

Figure S10: Spatial distribution of the monthly coefficient of variation (COV) for (a) MSA and (b) nss-$SO_4^=$ over the entire NA domain in 1998–2022. The COV for each grid point is calculated as the ratio between the standard deviation and the mean value, expressed in percentage (on log-transformed data), to evaluate the monthly stability of MSA and nss-$SO_4^=$. Higher COV indicates lower compound stability (many more variants).

Model Type	Preset	MSA				nss-$SO_4^{=}$			
		Cross-validation		Test		Cross-validation		Test	
		RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
Multilinear	Linear	0.408	0.74	0.392	0.76	0.324	0.53	0.318	0.55
Support Vector Machines	Linear	0.409	0.74	0.393	0.76	0.325	0.53	0.319	0.54
	Quadratic	0.398	0.75	0.384	0.77	0.318	0.55	0.310	0.57
	Cubic	0.390	0.76	0.376	0.78	0.311	0.57	0.304	0.59
	Fine Gaussian	0.410	0.74	0.399	0.75	0.299	0.60	0.293	0.61
	Medium Gaussian	0.382	0.77	0.370	0.79	0.305	0.58	0.296	0.61
	Coarse Gaussian	0.395	0.76	0.380	0.78	0.315	0.56	0.310	0.57
Decision Tree	Fine	0.457	0.67	0.457	0.67	0.357	0.43	0.355	0.43
	Medium	0.425	0.72	0.409	0.74	0.332	0.51	0.315	0.55
	Coarse	0.414	0.73	0.395	0.76	0.321	0.54	0.309	0.57
Regression Ensemble	Boosted	0.399	0.75	0.386	0.77	0.309	0.57	0.303	0.59
	Bagged	0.374	0.78	0.360	0.80	0.297	0.60	0.283	0.64
Gaussian Process Regression	Squared Exponential	0.383	0.77	0.370	0.79	0.305	0.58	0.299	0.60
	Matern 5/2	0.377	0.78	0.364	0.79	0.303	0.59	0.295	0.61
	Exponential	0.366	0.79	0.350	0.81	0.290	0.62	0.280	0.65
	Rational Quadratic	0.362	0.79	0.347	0.81	0.282	0.64	0.272	0.67
Neural Networks	Narrow	0.388	0.76	0.374	0.78	0.311	0.57	0.301	0.59
	Medium	0.387	0.77	0.379	0.78	0.311	0.57	0.300	0.60
	Wide	0.410	0.74	0.390	0.76	0.322	0.53	0.304	0.59
	Bi-layered	0.389	0.76	0.378	0.78	0.313	0.56	0.301	0.59
	Tri-layered	0.386	0.77	0.373	0.78	0.314	0.56	0.307	0.58

Table S1: Evaluation metrics for cross/validation and test datasets of the applied machine learning models and the multilinear model trained to estimate MSA and nss-$SO_4^{=}$ concentrations. Shaded cells represent the best performance from each type.

		MSA			nss-$SO_4^=$			MSA: nss-$SO_4^=$		
		Mean ± SD	Median	P10 P90	Mean ± SD	Median	P10 P90	Mean ± SD	Median	P10 P90
Annual		0.016 ± 0.007	0.017	0.007 0.023	0.250 ± 0.077	0.256	0.144 0.341	0.053 ± 0.012	0.054	0.034 0.063
Winter	Dec	0.006 ± 0.005	0.004	0.001 0.014	0.155 ± 0.079	0.138	0.066 0.273	0.032 ± 0.012	0.028	0.018 0.050
	Jan	0.006 ± 0.006	0.004	0.001 0.015	0.158 ± 0.086	0.131	0.064 0.288	0.032 ± 0.013	0.026	0.019 0.052
	Feb	0.008 ± 0.008	0.005	0.001 0.019	0.178 ± 0.101	0.140	0.067 0.319	0.036 ± 0.015	0.031	0.021 0.057
Spring	Mar	0.013 ± 0.010	0.011	0.002 0.023	0.219 ± 0.117	0.193	0.080 0.367	0.046 ± 0.018	0.047	0.026 0.065
	Apr	0.020 ± 0.012	0.023	0.003 0.032	0.279 ± 0.120	0.291	0.115 0.417	0.059 ± 0.020	0.067	0.029 0.078
	May	0.026 ± 0.014	0.028	0.006 0.042	0.337 ± 0.102	0.355	0.194 0.470	0.068 ± 0.022	0.075	0.030 0.094
Summer	Jun	0.029 ± 0.013	0.028	0.010 0.046	0.364 ± 0.075	0.358	0.280 0.474	0.074 ± 0.023	0.075	0.035 0.102
	Jul	0.029 ± 0.011	0.028	0.017 0.044	0.357 ± 0.060	0.352	0.293 0.442	0.077 ± 0.022	0.074	0.054 0.108
	Aug	0.025 ± 0.008	0.024	0.015 0.035	0.320 ± 0.059	0.323	0.248 0.395	0.072 ± 0.018	0.069	0.055 0.096
Autumn	Sep	0.016 ± 0.007	0.017	0.006 0.024	0.259 ± 0.080	0.286	0.128 0.346	0.056 ± 0.012	0.058	0.038 0.070
	Oct	0.010 ± 0.006	0.010	0.002 0.017	0.201 ± 0.086	0.219	0.078 0.301	0.042 ± 0.014	0.044	0.022 0.057
	Nov	0.007 ± 0.006	0.006	0.001 0.015	0.170 ± 0.084	0.175	0.066 0.280	0.035 ± 0.014	0.034	0.019 0.053

Table S2. Statistics of the annual and monthly climatology (1998-2022) of MSA, nss-$SO_4^=$ and MSA:nss-$SO_4^=$ (the maps shown in Fig.8 and Fig.9 (Main Text)). SD stands for spatial standard deviation. P10 and P90 represent the 10 and 90 percentiles, respectively.