Supplement to the article

Diagnosing sea-surface dimethylsulfide (DMS) concentration from satellite data at global and regional scales

Martí Galí[1], Maurice Levasseur[1], Emmanuel Devred[2], Rafel Simó[3], Marcel Babin[1]

[1] Takuvik Joint International Laboratory (Université Laval – CNRS). Biology Department, Université Laval. 1045 Avenue de la Médecine G1V 0A6 Québec (QC) Canada
[2] Fisheries and Oceans Canada, Bedford Institute of Oceanography, Dartmouth, NS B2Y 4A2, Canada
[3] Institut de Ciències del Mar (ICM-CSIC). Passeig Marítim de la Barceloneta 37-49, 08003 Barcelona, Catalonia, Spain

Correspondence to: Martí Galí (marti.gali.tapias@gmail.com)

This file contains detailed information regarding data sources and pre-processing of the extended DMS database (sections 1-3), statistical analyses regarding the development, optimization and validation of the DMS_{SAT} algorithm (section 4), and algorithm implementation (section 5).

S1 Sea surface DMS database and quality control

In situ concentrations of DMS, DMSPt and chlorophyll a (Chl), accompanied by ancillary data (bottom depth, temperature, salinity, wind speed), were downloaded from the public sea-surface DMS database (https://saga.pmel.noaa.gov/dms/) on 13 April 2017. This database (n = 47745 for DMS) was complemented with additional datasets obtained by the authors' teams (n = 403 for DMS; Table S1) as detailed by (Galí et al., 2015). Quality control involved the deletion of DMS and DMSPt measurements potentially affected by methodological issues according to the criteria described by (Galí et al., 2015). DMS concentrations lower than 0.1 nM (0.4% of data) or higher than 100 nM (0.3% of data) were also removed. Removal of DMS < 0.1 nM is justified because standard gas chromatography methods, with a detection limit of a few pmol S, would require sample volumes of >50 mL to resolve such low concentrations, whereas most studies analyzed DMS in smaller sample volumes (Bell et al., 2012). DMS concentrations >100 nM are seldom measured in seawater and, in certain datasets obtained before the 2000s, can certainly be attributed to methodological artifacts, i.e. sparging unfiltered samples that contained *Phaecoystis* sp. (del Valle et al., 2009) or other DMS-producing phytoplankton sensitive to mechanical strain (Wolfe et al., 2002). After selecting surface data (depth ≤ 10 m), samples taken on the same day and within a radius of 100 m were averaged. The final dataset had 41304, 3700 and 9182 measurements for DMS, DMSPt and Chl, respectively, with 3637 DMS-DMSPt and 8141 DMS-Chl pairs.

S2 Satellite matchup data

Daily and 8-day level 3-binned (L3BIN) data from the SeaWiFS and MODIS-Aqua sensors (9.28 and 4.64 km resolution, respectively) were matched to simultaneous in situ data from the DMS database (see Table S2). Matchups

were done using individual pixels and the average of 3x3 and 5x5 pixel boxes centered on the in situ measurement location using SeaDAS 6.4 (Galí et al., 2015). For both sensors, the percentage of valid satellite matchups was around 10% and 40% for daily and 8-day composites, respectively. Merged satellite variables were created in order to increase the amount of data available for statistical analyses, after observing that inconsistencies between the two satellite datasets were small compared to other sources of uncertainty. The merged Chl_{SAT}, $Kd490_{SAT}$, PIC_{SAT} and PAR_{SAT} variables were created by averaging SeaWiFS and MODIS-Aqua match-ups with a hierarchical search procedure, i.e. prioritizing daily data over 8-day data and single-pixel data over 3x3 and 5x5 pixel box means. The resulting satellite matchups originated in 51% of cases from "quasi-simultaneous" SeaWiFS and MODIS-Aqua retrievals. The remaining 49% of observations was divided evenly between the two sensors. Daily and 8-day gridded SST (4.6 km) from the AVHRR sensor was also matched to the in situ database.

S3 Binning of the extended sea surface DMS database

Statistical analyses were conducted using (i) non-binned data, (ii) data binned by month and 5x5 latitutde-longitude bins (M5x5), and (iii) data binned by month and the 56 Longhurst biogeochemical provinces (MLongh) (Longhurst, 2010). For binned data, bins with less than 3 data counts (M5x5) and 5 data counts (MLongh) were discarded (for being poorly documented) in order to increase the robustness of regression models. These cutoff values are rather arbitrary, but similar results were obtained with slightly larger cutoff values. The statistics of data bins and the amount of bins and individual DMS measurements discarded through this procedure are shown in Table S3, showing that the amount of individual data points discarded through the binning procedure was <2.5%. The mean (median) data counts per bin were 26.5 (10) for M5x5 binned data and 132.6 (57) for MLongh binned data.

S4 Algorithm coefficients: uncertainty and optimization

To assess the uncertainty in fitted eq. 2 coefficients, we used the bootstrap method to produce 10^5 sets of regression coefficients for eq. 2 using the MLongh binned dataset. Fig. S2 shows the nonrandom relationships among eq. 2 coefficients.

Regression-derived coefficients were further optimized for global and regional scales using a constrained nonlinear optimization approach developed for this study.

The optimal coefficients of eq. 1 were obtained by minimizing a cost function different from RMSE (which is by definition the *cost function* minimized by least-squares regression). The best model was obtained with a cost function J defined as:

$$J = RMSE + abs(1 - R^2) + abs(1 - Slope_{MA}) \qquad \text{(eq. S1)},$$

where $Slope_{MA}$ is the major axis regression between observed and predicted fields. This cost function rewards the model coefficients that predict DMS with R^2 and $Slope_{MA}$ closest to 1. The goodness-of-fit statistics used in eq. S1 were calculated in $\log_{10}$ space using the same MLongh binned dataset. To obtain realistic solutions, we constrained

the optimization to the 99% confidence intervals of the 10^5 bootstrapped regression coefficients shown in Fig. S2 (MLongh binned dataset). The resulting optimal model (eq. 2f) had higher DMSPt (β) and PAR (γ) coefficients and a smaller y-intercept than eq. 2e, and moved the modeled DMS concentration closer to the 1:1 agreement line without degrading neither RMSE nor R^2 (Table 2). The optimized model coefficients were validated using an independent dataset as described in section 3.1.3 of the main text (see Fig. 4).

The same approach was used to optimize the eq. 2 coefficients for the Bermuda Atlantic Time Series (BATS) site. In this case we used the 3 years of monthly measurements (upper mixed layer means) to obtain the regionally tuned coefficients (eq. 2h), which were not further validated using independent datasets given that they were only used to demonstrate the portability of the algorithm.

S5 Algorithm implementation: data sources and processing chain

The full DMS_{SAT} algorithm (Fig. 2) was implemented to produce (i) a monthly global DMS_{SAT} climatology based on SeaWiFS climatological 1997-2010 data; and (ii) regional time series with 8-day resolution for the period 2003-2016 using MODIS-Aqua data. In both cases we used reprocessing 2014.0. The data sources are summarized in Table S2.

Global DMS_{SAT} fields were computed using ocean color data from the SeaWiFS 1/12° gridded monthly climatology (1997-2010) in combination with the 1/2° gridded monthly MLD climatology from MIMOC. The input Chl_{SAT} product was either the band-ratio algorithm OC4-OCI (the current standard NASA Chl algorithm) or the semi-analytical GSM algorithm (Maritorena et al., 2002). The euphotic layer depth (Zeu) was computed as either the 1% penetration depth of 490 nm radiation (Zeu = 4.6/Kd490) or the semi-analytical Zeu from (Lee et al., 2007).

Regional DMS_{SAT} time series between 2003 and 2016 for latitudes >45°N were computed using daily MODIS-Aqua data (4.64 km) combined with the MIMOC MLD climatology (linearly interpolated onto the MODIS-Aqua grid and a daily period). Unlike the global implementation, which used Chl and Kd490 directly downloaded from the NASA Ocean Color website, in this case we used remote sensing reflectance spectra (Rrs) to compute bio-optical variables. Chl_{SAT} was computed using either the OC3 band-ratio algorithm (MODIS version of the OC algorithm (O'Reilly et al., 1998)) or the GSM algorithm (Maritorena et al., 2002). The diffuse attenuation coefficient at 488 nm, Kd488, which is nearly equivalent to the SeaWiFS Kd centered on 490 nm, was computed using the semi-analytical algorithm of (Lee et al., (2005) and used to estimate Zeu = 4.6/Kd488.

Since non-climatological satellite data frequently contain data gaps caused by cloudiness, we applied a binning and gap-filling procedure to obtain full coverage. First, we calculated $DMSPt_{SAT}$ (Galí et al., 2015) using daily 4.64 km MODIS-Aqua data (native L3bin resolution). Daily 4.64 km data were then averaged into 6x6 pixel boxes (27.84 km macropixels) and 8-day periods. The remaining gaps (10% pixels) were successively filled with 8-day (9%) and monthly (1%) climatologies of each variable. Finally, DMS_{SAT} was calculated from 8-day 27.84 km $DMSPt_{SAT}$ and PAR_{SAT}.

We performed a further sensitivity test to analyze the effect of different mixed layer depth (MLD) products on $DMSPt_{SAT}$. Besides our standard MLD obtained from the MIMOC climatology (Schmidtko et al., 2013), we tested the monthly MLD time series diagnosed by the Global Ocean Data Assimilation System (GODAS) ocean circulation model (which covers latitudes <65° between 1980 and present). Eight-day $DMSPt_{SAT}$ time series between 2003 and 2015 were generated for two areas located in the North Atlantic (Iceland Basin) and Pacific (Bering Sea), which show contrasting wintertime MLD —owing to differences in salinity stratification. To verify the accuracy of MIMOC and GODAS MLD, we compared them to MLD derived from collocated ARGO float profiles. As shown in Fig. S1, GODAS overestimated wintertime MLD in the subpolar Atlantic. This translated, in some years, in slightly lower $DMSPt_{SAT}$ with GODAS because the DMSPt sub-algorithm switched at a later date from the 'mixed' (Zeu/MLD < 1) to the 'stratified' (Zeu/MLD > 1) waters equation. Differences were almost absent in the Bering Sea. Overall, the use of climatological or model-derived MLD had a negligible effect on diagnosed $DMSPt_{SAT}$.

Tables

Table S1. Compilation of studies added to the sea surface DMS database used for algorithm development and validation.

Reference	Region	Dates
Levasseur et al. 2006	NW Pacific	Jul 2002
Matrai et al. 2007	Barents Sea	1998, 1999, 2001
Royer et al. 2010	NW Pacific	Jul 2007
Lizotte et al. 2012, SABINA cruises	NW Atlantic	2003
Luce et al. (2011), IPY C-SOLAS cruises	Canadian Arctic	Late summer 2007, 2008
Malaspina circumnavigation cruise (Royer et al. 2015)	Tropical Atlantic, Pacific and Indian oceans	Dec 2010—Jul 2011
SUMMER cruises (Royer et al. 2016)	NW Mediterranean	Sep 2011, May 2012

Table S2. Summary of datasets used to complement the sea surface DMS database (for algorithm development and validation) and to implement the DMS_{SAT} algorithm at global and regional scales. R: reprocessing.

Data type and source	Use, data version and type (when applicable).
SeaWiFS and MODIS-Aqua ocean color data: https://oceancolor.gsfc.nasa.gov/	**Matchups**: SeaWiFS R2010.0 and MODIS-Aqua R2013.1. DAY and 8D L3BIN data. **Global** ocean implementation of DMS_{SAT}, SD02 and VS07 algorithms: MODIS-Aqua R2014.0. MONTH L3SMI data. **Regional** DMS_{SAT} implementation: MODIS-Aqua R2014.0. DAY L3SMI data. *Note: MODIS-Aqua nighttime SST was used between 2003-2016 instead of AVHRR SST*.
AVHRR sea surface temperature (SST): https://www.nodc.noaa.gov/SatelliteData/pathfinder4km/	**Matchups**: Pathfinder v5.2. **Global** DMS_{SAT} implementation: 1/12 degree monthly climatology based on Pathfinder v5.2. **Regional** DMS_{SAT} implementations of DMS_{SAT}: Pathfinder v5.3. *Note: MODIS-Aqua nighttime SST was used between 2003-2016 instead of AVHRR SST*.
Monthly Isopycnal & Mixed-layer Ocean Climatology (MIMOC): http://www.pmel.noaa.gov/mimoc/	**Matchups**: Native 0.5 degrees monthly resolution. **Global** ocean implementation of DMS_{SAT}, SD02 and VS07 algorithms: Native 0.5 degrees monthly resolution. **Regional** DMS_{SAT} implementation: Reprojected onto MODIS-Aqua 4.64 sinusoidal grid and interpolated to 1-day resolution.
NCEP Global Ocean Data Assimilation System (GODAS): https://www.esrl.noaa.gov/psd/data/gridded/data.godas.html	**Regional** DMS_{SAT} implementation sensitivity tests: Reprojected onto MODIS-Aqua 4.64 sinusoidal grid and interpolated to 1-day resolution.
ARGO float MLD: http://mixedlayer.ucsd.edu/	**Regional** DMS_{SAT} implementation sensitivity tests: matched to 27.84 8-day resolution of DMS_{SAT} dataset.

World Ocean Atlas 2009 nutrient and salinity data: http://coastwatch.pfeg.noaa.gov/erddap/griddap/nodcWoa09sea1n.html	**Matchups**: 1-degree monthly resolution.
General Bathymetric Chart of the Oceans https://www.gebco.net/data_and_products/gridded_bathymetry_data/gebco_one_minute_grid/	**Matchups:** GEBCO One Minute Grid (November 2008)

Table S3. Binning statistics for the extended sea surface DMS database. V = valid, D = discarded.

Binning scheme	**Minimum N per bin**	**Bin counts**			**Individual data counts**		
		V	D	% D	V	D	% D
M5x5	$N \geq 3$	1562	742	32.2%	40326	978	2.4%
Mlongh	$N \geq 5$	322	66	17.0%	41156	148	0.4%

Table S4. Summary of the stepwise regression analysis. Only models with significant coefficients are shown. WS: wind speed.

Model	**R^2adj**	**RMSE**	**AIC**	**N**
Non-binned data				
$\log_{10}$DMS = -1.21 + 0.67 $\log_{10}$DMSPt + 0.0136 PAR	0.50	0.35	2743	3620
$\log_{10}$DMS = -1.01 + 0.75 $\log_{10}$DMSPt + 0.0111 PAR_{MLD}	0.47	0.37	2992	3595
$\log_{10}$DMS = -1.17 + 0.64 $\log_{10}$DMSPt + 0.0151 PAR - 0.0038 SST	0.51	0.35	2722	3615
$\log_{10}$DMS = -1.22 + 0.63 $\log_{10}$DMSPt + 0.0152 PAR + 0.055 $\log_{10}[NO_3]$	0.51	0.34	2357	3418
$\log_{10}$DMS = -1.17 + 0.64 $\log_{10}$DMSPt + 0.0147 PAR - 0.040 $\log_{10}$[N-cline]	0.51	0.35	2562	3418
$\log_{10}$DMS = -1.79 + 0.70 $\log_{10}$DMSPt + 0.0159 PAR + 0.012 Salinity	0.52	0.31	1027	1911
$\log_{10}$DMS = -1.04 + 0.59 $\log_{10}$DMSPt + 0.0118 PAR - 0.0082 WS	0.48	0.32	764	1442
$\log_{10}$DMS = -0.63 + 0.45 $\log_{10}$DMSPt + 0.0129 PAR + 0.098 $\log_{10}PIC_{SAT}$	0.35	0.29	388	1123
MLongh binned data (bin medians)				
$\log_{10}$DMS = -1.02 + 0.45 $\log_{10}$DMSPt + 0.0163 PAR	0.57	0.21	-31.0	118
$\log_{10}$DMS = -0.94 + 0.69 $\log_{10}$DMSPt + 0.0172 PAR_{MLD}	0.52	0.25	15.7	118
$\log_{10}$DMS = -0.91 + 0.48 $\log_{10}$DMSPt + 0.0189 PAR - 0.0087 SST	0.59	0.24	0.9	118
$[NO_3]$ coefficient non-significant (p = 0.26)				118
[N-cline] coefficient non-significant (p = 0.71)				118
Salinity coefficient non-significant (p = 0.35)				102
WS coefficient non-significant (p = 0.86)				97
$\log_{10}$DMS = -0.64 + 0.29 $\log_{10}$DMSPt + 0.0118 PAR + 0.033 $\log_{10}PIC_{SAT}$	0.52	0.21	-14.8	86

Table S5. DMS_{SAT} validation statistics for constrained and *unconstrained (in italics)* Chl_{SAT} error. N increases from 86 to 1293 as the tolerated Chl_{SAT} error increases. N is 14677 for unconstrained Chl error.

Algorithm	**R^2 (log_{10} space)**	**RMSE (log_{10} space)**	**MAPE (linear space)**	**Mean bias (linear space)**
DMS_{SAT} (eq. 2f)	0.40–0.52 *0.29*	0.21–0.30 *0.38*	42–68% *108%*	7–29% *-9%*
SD02 (eq. 3)	0.22–0.30 *0.24*	0.21–0.31 *0.40*	42–78% *138%*	9–31% *2%*
VS07 (eq. 4)	0.10–0.22 *0.02*	0.30–0.34 *0.45*	82–94% *89%*	0–42% *-41%*

Figures

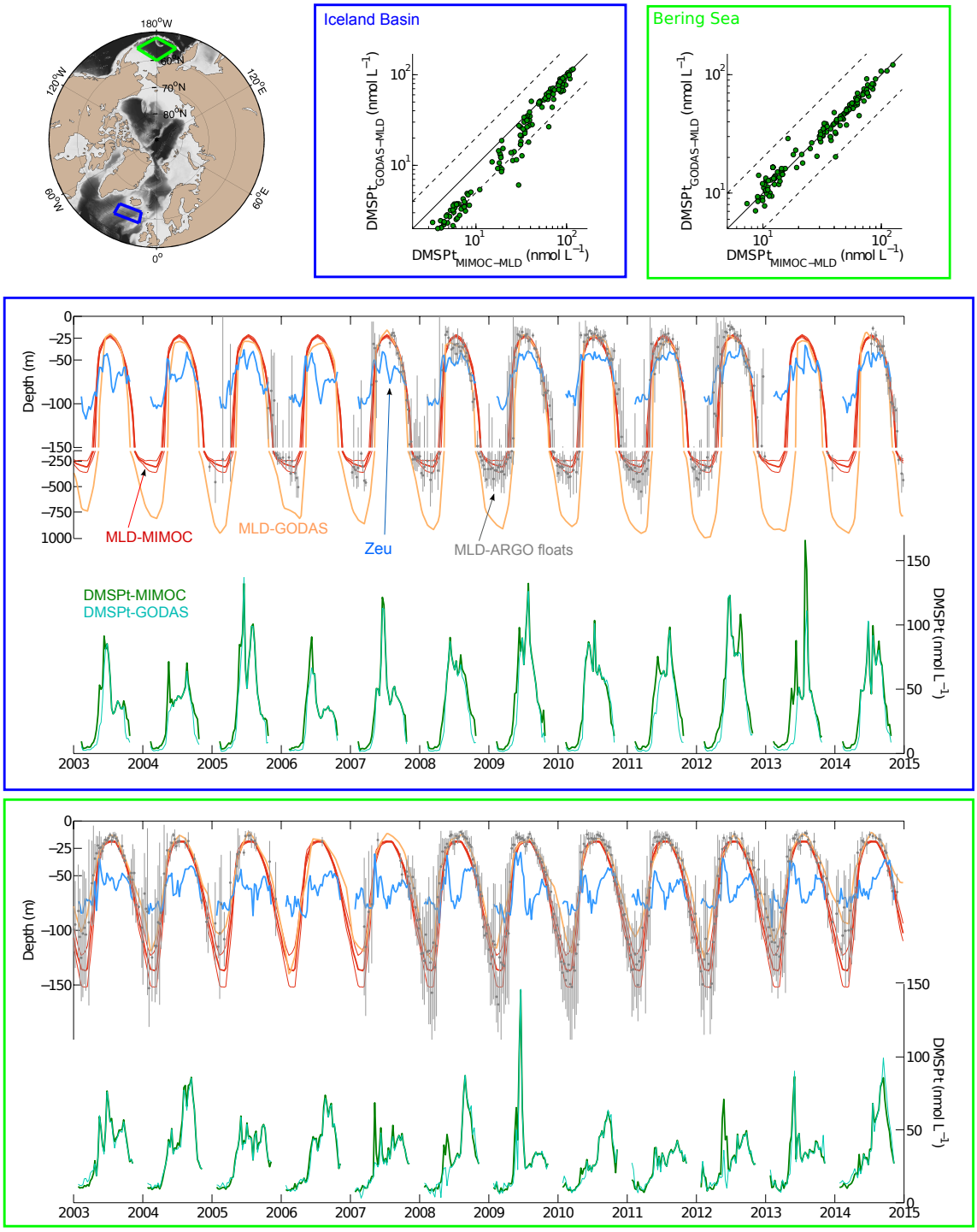

Figure S1: Sensitivity of the DMSPt$_{SAT}$ sub-algorithm to different MLD input data. Top: scatterplots comparing DMSPt$_{SAT}$ calculated with the MIMOC and GODAS MLD products. Middle and bottom: 2003-2015 8-day resolution time series of MIMOC and GODAS MLD, ARGO float MLD, and satellite derived euphotic layer depth (Zeu). The scatterplots and time series plots show the regional means of each variable for the two regions highlighted in the map: Iceland Basin and Bering Sea.

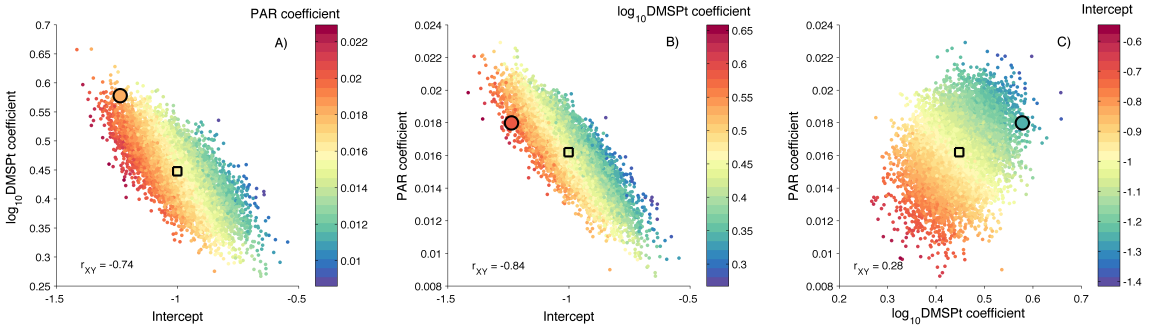

Figure S2. Bootstrapped regression model coefficients obtained for the equation: $\log_{10}(DMS) = \alpha + \beta \log_{10}(DMSPt) + \gamma PAR$, with n = 10000. Black squares show the mean coefficients, which are equivalent to those obtained through regular multiple regression (eq. 2e). The uncertainty envelopes defined by the 10000 bootstrapped coefficients were used as the bounds for a constrained optimization procedure (see text). Filled circles represent the resulting optimized coefficients (eq. 2f).

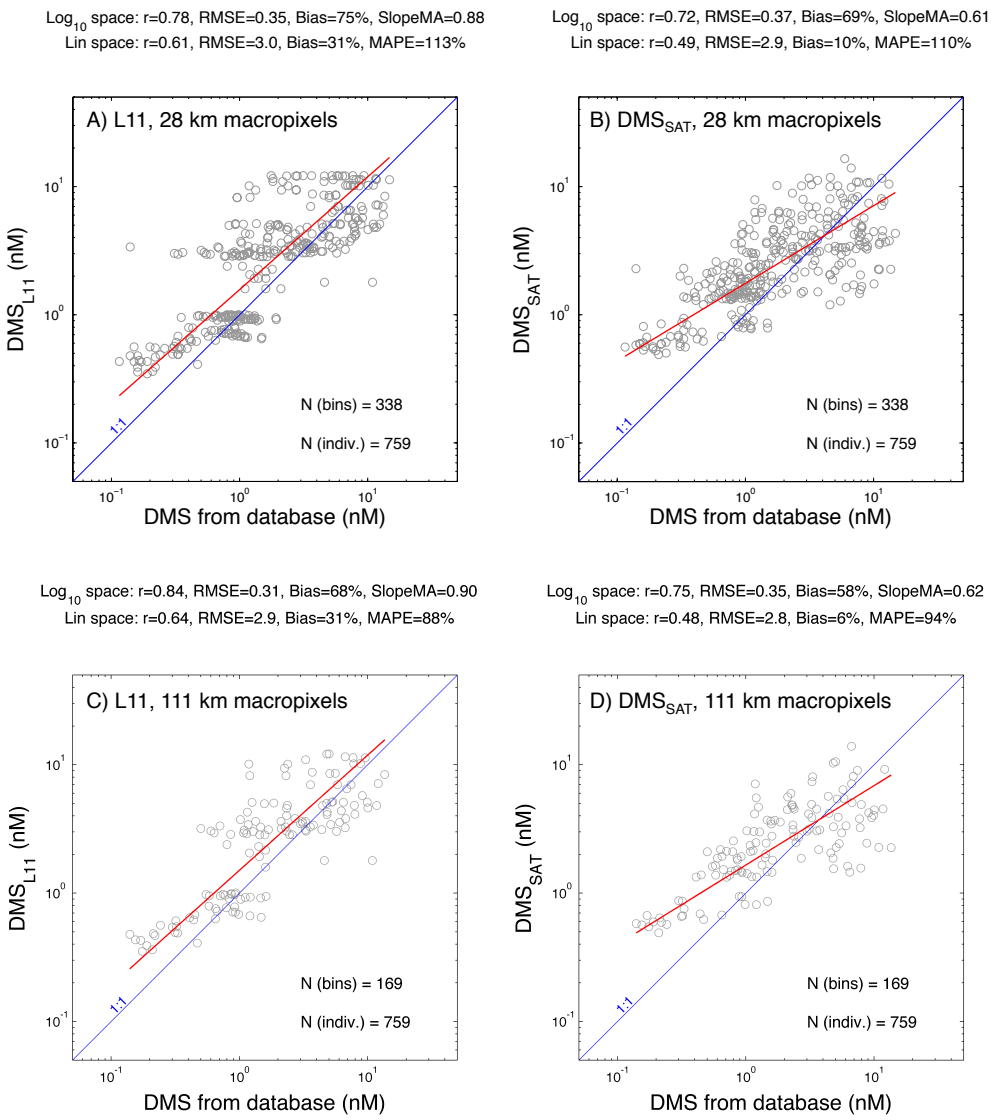

Figure S3. **DMS_{SAT} validation statistics for the subpolar North Atlantic**. In situ data points correspond to the area delimited by latitudes between 45°N–60°N and longitudes between 55°W–15°E, marked in Fig. 9 of the paper. Left-hand plots compare the L11 climatology to the in situ measurements (on which it is based). Right-hand plots compare DMS_{SAT} to the same in situ measurements. Top and bottom plots differ in the degree of spatial binning applied: 28 km and 111 km. The temporal binning period is 8 days in all cases. On top of each plot we provide both $\log_{10}$ and linear space statistics. The perfect agreement line is shown in blue and the model-data linear regression in red (major axis regression, $\log_{10}$ space).