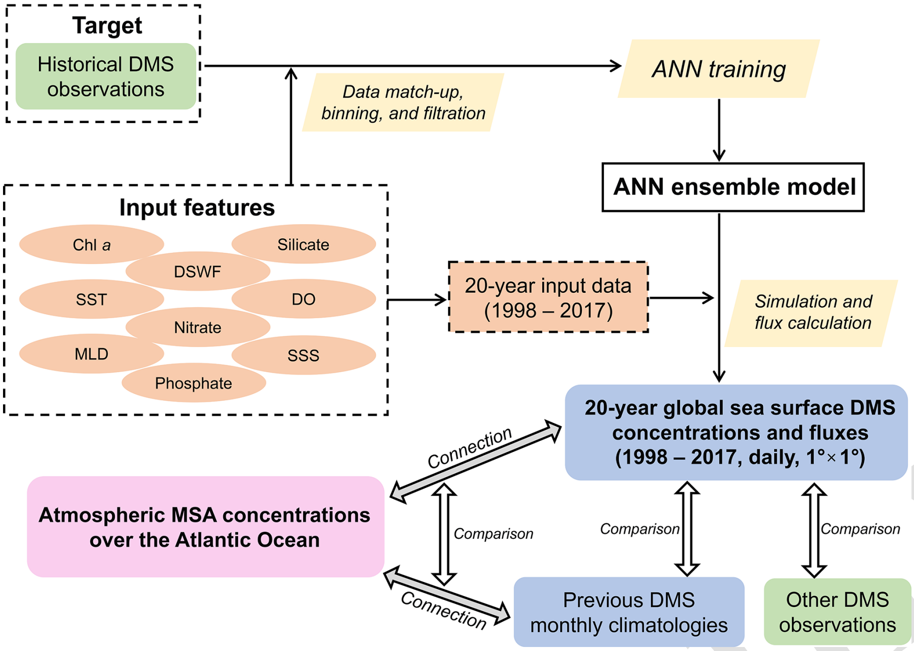
Target
Historical DMS observations
Data match-up, binning, and filtration
ANN training
ANN ensemble model
Input features
Chl a
Silicate
DSWF
SST
DO
Nitrate
MLD
SSS
Phosphate
20-year input data (1998 – 2017)
Simulation and flux calculation
20-year global sea surface DMS concentrations and fluxes (1998 – 2017, daily, 1°×1°)
Connection
Atmospheric MSA concentrations over the Atlantic Ocean
Comparison
Comparison
Comparison
Connection
Previous DMS monthly climatologies
Other DMS observations

Figure 1. Flowchart of this study, including the development of the ANN ensemble model, the construction of the new DMS gridded dataset, and subsequent evaluations of this product.

try and high-performance liquid chromatography (HPLC). In order to improve mutual consistency, a conversion between the data from these two methods was applied, and then the in situ Chl *a* concentrations were adjusted to match them up with the satellite Chl *a* using the functions described in Galí et al. (2015). After that, the statistical outliers among all the $\log_{10}$(Chl *a*) values (i.e., those outside a range defined as the average ±3 standard deviations) were eliminated. A comparison between the in situ and satellite-retrieved Chl *a* data is shown in Fig. S2. The strong consistency between in situ and daily satellite Chl *a* data ($R^2 > 0.5$; RMSE < 0.4) provides the rationale for integrating these datasets. The $\log_{10}$ transformation was applied to make the data distribution close to a normal distribution. When finally selecting the $\log_{10}$(Chl *a*) value corresponding to each DMS data value, in situ data were prioritized where available; otherwise, the satellite-retrieved data were used.

The DMS values and extracted values of MLD and three nutrients (nitrate, phosphate, and silicate) were also $\log_{10}$-transformed. The statistical outliers for each variable were excluded as mentioned above. After data filtration, a total of 633 361CE1 samples with valid data for all variables were obtained. To avoid a data aggregation bias stemming from the clustering of multiple data points within a narrow temporal range and spatial range (i.e., obtained on the same day and within a region smaller than 0.05° × 0.05°), these data points were averaged. Consequently, 41 157 binned samples were utilized for subsequent model development; their spatial distribution is depicted in Fig. 2a.

We divided the global ocean into nine regions based on Longhurst's biomes (Longhurst, 1998). There are six biomes in Longhurst's definition, including Coastal, Polar_N, Polar_S, Westerlies_N, Westerlies_S, and Trades (the .shp file of Longhurst's biomes and provinces was downloaded from https://www.marineregions.org/downloads.php#longhurst, last access: 16 April 2020). We further divided Westerlies_N into Westerlies_N_Pacific and Westerlies_N_Atlantic and divided Trades into Trades_ Pacific, Trades_Indian, and Trades_Atlantic based on the different oceanic basins, as shown in Fig. 2b. It is noteworthy that there are 11 237 samples in the Coastal region, constituting 27.3 % of the entire sample set, despite the Coastal biome accounting for only 9.7 % of the global ocean area. Given the distinct physiochemical and biological conditions of seawater in coastal seas compared to other regions, the disproportionately higher density of samples within the Coastal biome might cause the model to overly prioritize this region. To mitigate this data imbalance and ensure the model captures broader patterns in open oceans, we adjusted the data distribution during the model training and validation processes. Specifically, for each training session, a portion of coastal samples was randomly removed to ensure that the proportion of coastal samples in the total sample set (denoted as F_{coastal}) matched the coastal proportion of the total area.

2.3 Artificial neural network training and validation

The 41 157 binned samples obtained after the previously mentioned data preprocessing were used to develop the ar-