Supplement of Biogeosciences, 22, 5975–5990, 2025
https://doi.org/10.5194/bg-22-5975-2025-supplement

Biogeosciences
Open Access

Supplement of

Improving marine sediment carbon stock estimates: the role of dry bulk density and predictor adjustments

Mark Chatting et al.

Correspondence to: Mark Chatting (mark.chatting@ucd.ie)

Table S1: Summary of sources of legacy organic carbon data.

Source	Number of Data points	Access	Link
Agri-Food & Biosciences Institute Northern Ireland (AFBI)	93	Private	
Archived samples (UCD/DCU)	45	Private	
Natural Resources Wales	70	Public with request	https://naturalresources.wales/evidence-and-data/accessing-our-data/request-environmental-data/?lang=en
International Council for the Exploration of the Sea (ICES)	931	Public	https://datras.ices.dk/Home/Access.aspx
INFOMAR	20	Public with request	https://www.marine.ie/data-request
Mason et al. (2017)	137	Public	https://data.cefas.co.uk/view/18354
MERC consultancy	12	Private	
Marine Institute	235	Public with request	https://www.marine.ie/data-request
Natura PSA	127	Public with request	https://www.marine.ie/data-request

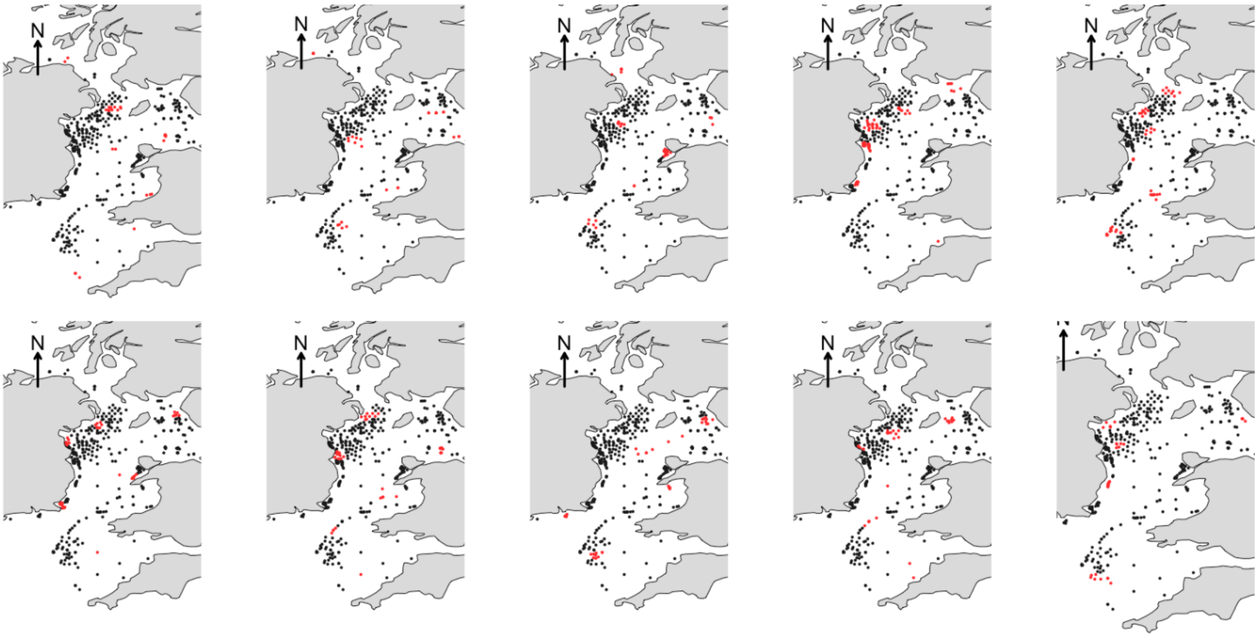

Fig. S1: Spatial folds showing training/test data splits. Points used to train the $OC_{content}$ Random Forest model are coloured black, while red data points were used to test performance for each fold. The k-Nearest Neighbour Distance Matching (kNNDM) function was used to ensure spatial independence between training and testing splits.

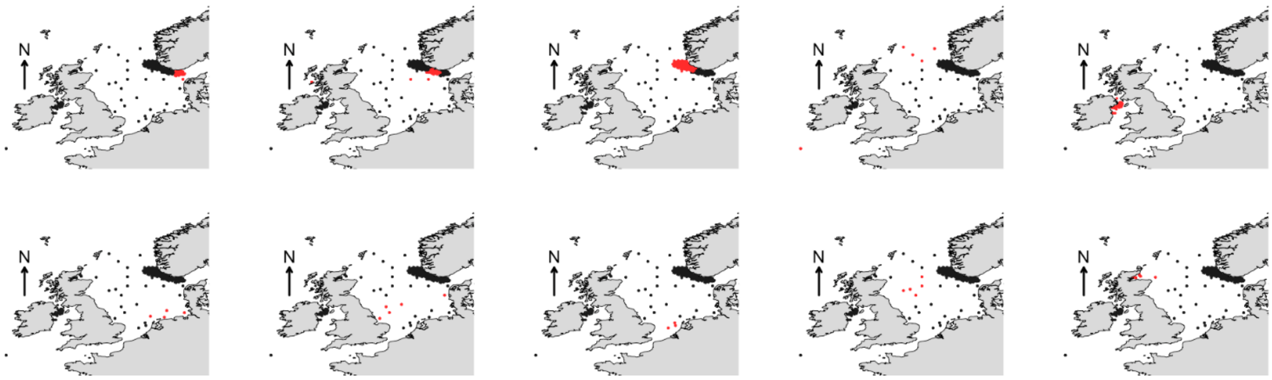

Fig. S2: Spatial folds showing training/test data splits. Points used to train the DBD Random Forest model are coloured black, while red data points were used to test performance for each fold. The k-Nearest Neighbour Distance Matching (kNNDM) function was used to ensure spatial independence between training and testing splits.

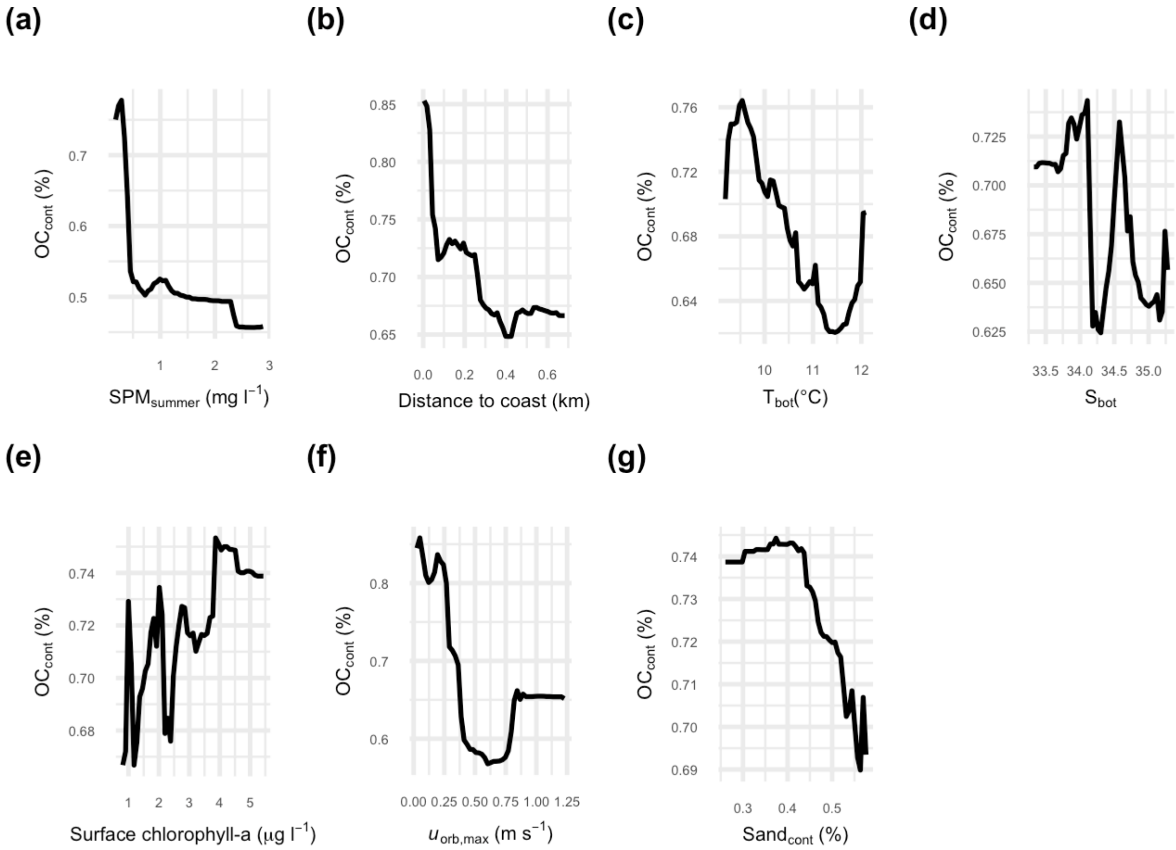

Fig. S3: Partial dependence plots showing the relationship between OC content and non-bias adjusted model predictors selected by Forward Feature Selection (FFS): (a) surface summer suspended particulate matter, (b) distance to the nearest coast, (c) bottom water temperature, (d) bottom water salinity, (e) surface chlorophyll-a, (f) maximum wave orbital velocity at the seafloor and (g) sand content.

Table S2: Summary of different model predictors importance on Mean Squared Error (MSE) for the three Random Forest models trained ($OC_{cont,pre}$, $OC_{cont,post}$ and DBD_{post}).

Predictor	**% Increase in model MSE**
$OC_{cont,pre}$	
SPM_{summer}	37.1
Distance to the nearest coast	23.9
T_{bot}	23.4
S_{bot}	22.4
Chlorophyll-a	21.2
$u_{orb,max}$	20.6
$Sand_{cont}$	15.8
$OC_{cont,post}$	
mud_{cont}	56.8
$u_{orb,max}$	32.4
Distance to the nearest coast	23.9
Chlorophyll-a	23.8
Bathymetry	21.7
DBD_{post}	
$Sand_{cont}$	45.9
SPM_{summer}	32.2
SPM_{winter}	26.3
$u_{orb,mean}$	22.5
$u_{orb,max}$	21.9
$U_{bot,mean}$	19.6

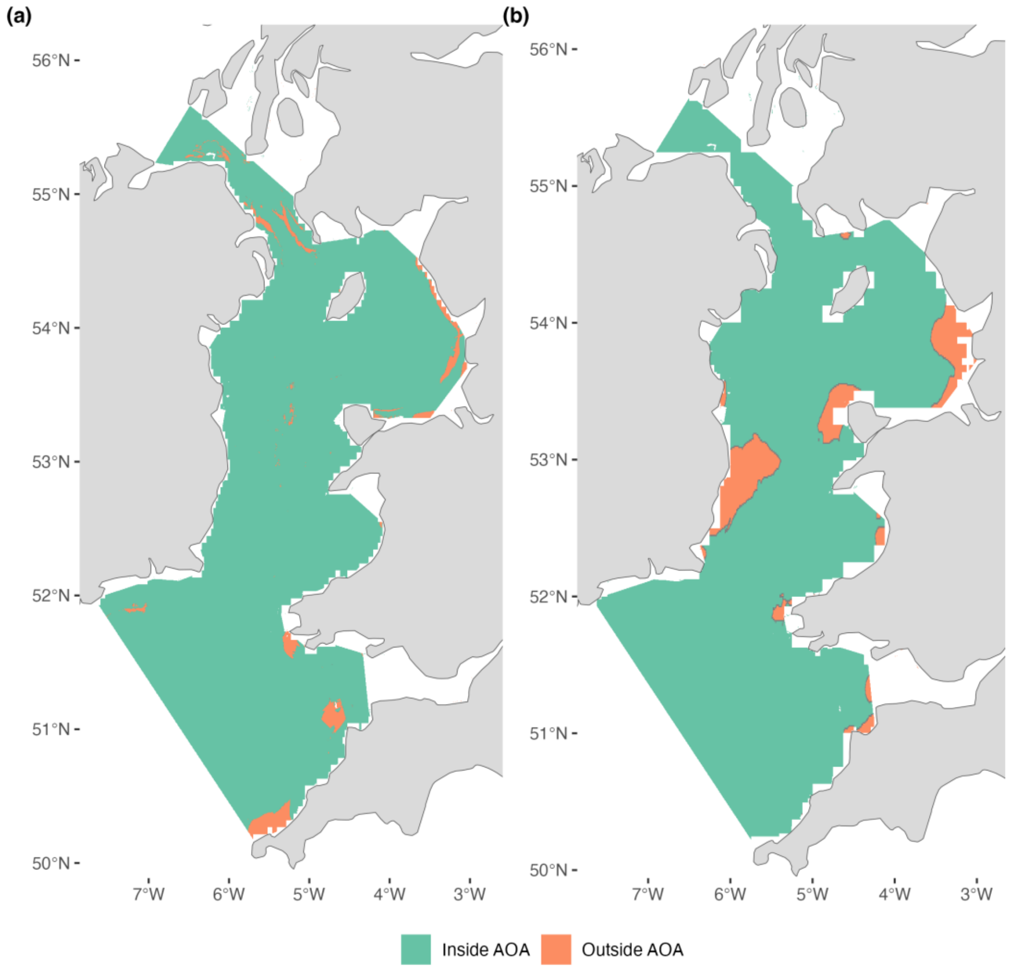

Fig. S4: Area of Applicability (AOA) for (a) adjusted OC content and (b) adjusted DBD. Pixels that are labelled 'inside AOA' represent locations where training and prediction data are within a determined dissimilarity threshold and are similar enough to assume a robust model prediction.

Supplementary methods

Predictor data

Distance to coast: Distance to coast was calculated using the distance() function in the terra package (https://cran.r-project.org/web/packages/terra/index.html) in R. This function calculates the distance to a predefined coastline in metres using longitude and latitude, which is unlike Mitchell et al. (2019) who calculated Euclidean distances.

Bathymetry: A bathymetric Digital Terrain Model (DTM) was downloaded from the EMODnet-bathymetry portal (http://www.emodnet-bathymetry.eu/) (EMODNet REF) and was clipped to the study area. EMODnet-bathymetry data has a grid size of approximately 155 m by 230 m. As this was the finest resolution predictor data, all other predictor variables (see below) were resampled to this same grid size using the resample() function in the terra package in R.

Predictors prior to adjustment (Predictors$_{pre}$)

Bottom water salinity, bottom water temperature, mean and maximum bottom water velocities, surface chlorophyll-a, surface suspended particulate matter (winter and summer): Bottom water temperature (BWT), bottom water salinity (BWS), mean and maximum bottom water velocities surface chlorophyll-a and surface suspended particulate matter were all obtained from the Copernicus marine data portal (https://data.marine.copernicus.eu/products). The Global Ocean Physics Reanalysis data set (GLOBAL_MULTIYEAR_PHY_001_030) was used for BWT, BWS, and mean and maximum bottom water velocities. Surface chlorophyll-a, summer surface suspended particulate matter (SPM) winter SPM were all obtained from the Global Ocean Colour from satellite observations (OCEANCOLOUR_GLO_BGC_L4_MY_009_104) data set. Monthly averages between January 2000 and December 2019 were downloaded at a 4 km resolution for each of these predictors. In order to produce summer and winter averages for surface suspended particulate matter, data were averaged across the 20 years for the summer months (June, July and August) and winter months (December, January and February).

Sediment properties (mud, sand and gravel content): Recent research has produced several versions of sediment property datasets for the NW European Shelf (Mitchell et al. 2019; Stephens & Diesing 2015; Wilson et al. 2018). Given the significant influence of sediment properties on OC stock predictions (Diesing et al. 2017; Smeaton et al. 2021), it was critical to empirically justify the selection of mud, sand, and gravel content used to train $OC_{content\ pre}$ to avoid artificially reducing its performance relative to $OC_{content\ post}$. To achieve this, observation data from the Marine Institute (https://erddap.marine.ie/erddap/tabledap/IMI_CTD.html) and Mitchell et al. (2019) were used to calculate the RMSE for the models of Mitchell et al. (2019), Stevens & Diesing (2015), and Wilson et al. (2018). Additionally, as sediment data are compositional, bounded by 0 and 1 and must sum to 1 these data were pre-treated prior to averaging. Additive log ratio (ALR) transformations were applied prior to averaging using Eq. 2 and Eq. 3 (Mitchell et al., 2019):

$$ALR_m = \log\left(\frac{mud}{gravel}\right), \quad \text{(Eq. S1)}$$

$$ALR_s = \log\left(\frac{sand}{gravel}\right), \quad \text{(Eq. S2)}$$

After averaging ALR transformed values, (Mitchell et al., 2019; Stephens and Diesing, 2015; Wilson et al., 2018)they were back transformed to compositional data using the following Eq. 4, Eq. 5 and Eq. 6 (Mitchell et al., 2019):

$$mud = \frac{\exp(ALR_m)}{\exp(ALR_m)+\exp(ALR_s)+1}, \quad \text{(Eq. S3)}$$

$$sand = \frac{\exp(ALR_s)}{\exp(ALR_s)+\exp(ALR_m)+1}, \quad \text{(Eq. S4)}$$

$$gravel = 1 - (mud + sand), \quad \text{(Eq. S5)}$$

These back transformed values were used as the final mud, sand and gravel predictors in $\text{predictors}_{\text{post}}$. Maximum and mean wave orbital velocity at the seafloor: Mean and maximum wave orbital velocities at the seafloor were sourced from Wilson et al. (2018). Wave conditions, derived from the ERA-Interim reanalysis (Dee et al., 2011), were integrated with significant wave height, mean wave period, and mean wave direction (ECMWF: http://www.ecmwf.int/en/research/climate-reanalysis/). The ERA-Interim reanalysis features a spatial resolution of approximately 79 km and a temporal resolution of 6 hours.

Observational data used to bias adjust inputs

Data repositories Pangaea (https://www.pangaea.de) and the Marine Institute (https://erddap.marine.ie/erddap/tabledap/IMI_CTD.html) were searched for in situ measurement data. When extracting bottom water observational data, EMODNet bathymetry was used to ensure data points were <2m from the sea floor. If surface predictors were being extracted, only data points from <2m water depth were used. As $\text{predictors}_{\text{pre}}$ data were subset to between the years 2000 and 2020 observational data was constrained to between these years.

Quantile-quantile mapping bias adjustment

As in situ data represents a measurement taken at one point in space and time, they were smoothed both in space and in time in order to be harmonized with model data which represents monthly averages (Cheng et al. 2017; Cheng et al. 2020). Point grids of each month were created by grouping data by month applying a nine-point filter to each data point for each month. Then, 3 months of data were merged together centred around the month of interest. For example, if creating a smoothed point grid for June, point data from May, June and July were merged together. Smoothed point grids were interpolated using Inverse Distance Weighting (IDW) to create a continuous surface from point data for June. This process was repeated for each month, resulting in 12 interpolated surfaces for each month of the year. All 12 monthly surfaces were averaged together to obtain the annual average. This annual average continuous surface was then used to bias adjust the $predictors_{pre}$ using QQ mapping, resulting in a model that has been bias adjusted based on local in situ measurements (Cheng et al. 2017; Cheng et al. 2020).

Supplementary Information References

Cheng, L., Trenberth, K. E., Fasullo, J., Boyer, T., Abraham, J., & Zhu, J. (2017). Improved estimates of ocean heat content from 1960 to 2015. *Science Advances*, *3*(3). https://doi.org/10.1126/sciadv.1601545

Cheng, L., Trenberth, K. E., Gruber, N., Abraham, J. P., Fasullo, J. T., Li, G., Mann, M. E., Zhao, X., & Zhu, J. (2020). Improved Estimates of Changes in Upper Ocean Salinity and the Hydrological Cycle. *Journal of Climate*, *33*(23), 10357–10381. https://doi.org/10.1175/JCLI-D-20-0366.1

Haimberger, L. and Tavolato, C., 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society: a journal of the atmospheric sciences, applied meteorology, and physical oceanography*, *137*(656), pp.553-597.

Diesing, M., Kröger, S., Parker, R., Jenkins, C., Mason, C., & Weston, K. (2017). Predicting the standing stock of organic carbon in surface sediments of the North–West European continental shelf. *Biogeochemistry*, *135*(1–2), 183–200. https://doi.org/10.1007/s10533-017-0310-4

Mitchell, P. J., Aldridge, J., & Diesing, M. (2019). Legacy Data: How Decades of Seabed Sampling Can Produce Robust Predictions and Versatile Products. *Geosciences*, *9*(4), 182. https://doi.org/10.3390/geosciences9040182

Smeaton, C., Hunt, C. A., Turrell, W. R., & Austin, W. E. N. (2021). Marine Sedimentary Carbon Stocks of the United Kingdom's Exclusive Economic Zone. *Frontiers in Earth Science*, *9*. https://doi.org/10.3389/feart.2021.593324