1 **Predictive mapping of organic carbon stocks in surficial sediments of**

2 **the Canadian continental margin**

4 Graham Epstein[1*], Susanna D. Fuller[2], Dipti Hingmire[3], Paul G. Myers[4], Angelica Peña[5], Clark Pennelly[4]

5 & Julia K. Baum[1]

7 [1]Department of Biological Sciences, University of Victoria, Victoria, British Columbia, Canada, V8P 5C2

8 [2]Oceans North, Halifax, NS B3J 1E6, Canada.

9 [3]School of Earth and Ocean Sciences (SEOS), University of Victoria, Victoria, British Columbia, Canada,

10 V8P 5C2

11 [4] Department of Earth and Atmospheric Sciences, University of Alberta, Edmonton, Canada, AB T6G 2E3

12 [5] Institute of Ocean Sciences, Fisheries and Ocean Canada, Sidney, British Columbia, Canada, V8L 4B2

14 *Corresponding author: Email - grahamepstein@uvic.ca

Abstract

The quantification and mapping of surficial seabed sediment organic carbon has wide-scale relevance across marine ecology, geology and environmental resource management, with carbon densities and accumulation rates being a major indicator of geological history, ecological function, and ecosystem service provisioning, including the potential to contribute to nature-based climate change mitigation. While global analyses can appear to provide a definitive understanding of the spatial distribution of sediment carbon, regional maps may be constructed at finer resolutions and can utilise targeted data syntheses and refined spatial data products and therefore have the potential to improve these estimates. Here, we report a national systematic review of data on organic carbon content in seabed sediments across Canada and combine this with a synthesis and unification of best available data on sediment composition, seafloor morphology, hydrology, chemistry and geographic settings within a machine learning mapping framework. Predictive quantitative maps of mud content, dry bulk density, organic carbon content and organic carbon density, were produced along with cell specific estimates of their uncertainty at 200 m resolution across 4,489,235 km^2 of the Canadian continental margin (92.6% of the seafloor area above 2,500 m). Fine-scale variation in carbon stocks was identified across the Canadian continental margin, particularly in the Pacific and Atlantic Ocean regions. Overall, we estimate the standing stock of organic carbon in the top 30 cm of surficial seabed sediments across the Canadian shelf and slope to be 10.9 Gt (7.0 – 16.0 Gt). Increased empirical sediment data collection and higher precision in spatial environmental data-layers could significantly reduce uncertainty and increase accuracy in these products over time.

1. Introduction

The organic carbon contained in seafloor sediments has a major influence on global carbon cycles and earth's climate (Hülse et al., 2017; Bauer et al., 2013). Seabed sediments have been estimated to accumulate approximately 126–350 Mt of organic carbon per year (Keil, 2017; Berner, 1982) and contain 87 Gt of organic carbon in their top 5 cm (Lee et al., 2019), 168 Gt in the top 10 cm (LaRowe et al., 2020a) and up to ~2,300 Gt in the top 1 m (Atwood et al., 2020), with the latter being equivalent to nearly twice that of soils on land. Continental shelves have the highest densities of sediment carbon across the global ocean, covering only 5-8% of the marine area but an estimated 15-19% of surficial organic carbon stocks (LaRowe et al., 2020a; Atwood et al., 2020) and 80% of annual carbon burial (Bauer et al., 2013; Burdige, 2007). Continental

48   margin zones (continental shelves and slopes) also contain the largest spatial variation in organic
49   carbon due to highly heterogenous geological, geographic, biological and oceanographic settings
50   (Smeaton et al., 2021; Diesing et al., 2017, 2021; Atwood et al., 2020). They are also subjected
51   to high levels of human activity, being impacted by many coastal and marine industries including
52   fishing, shipping, energy generation, telecommunication, mineral extraction, and pollution from
53   land based activities (Halpern et al., 2019; Amoroso et al., 2018; Keil, 2017). The quantification
54   and mapping of organic carbon on continental margins is therefore imperative for best practise
55   seabed management; with the densities and accumulation rates being a major indicator of
56   ecological function, geological history and ecosystem service provision (Legge et al., 2020;
57   Snelgrove et al., 2018; Middelburg, 2018).

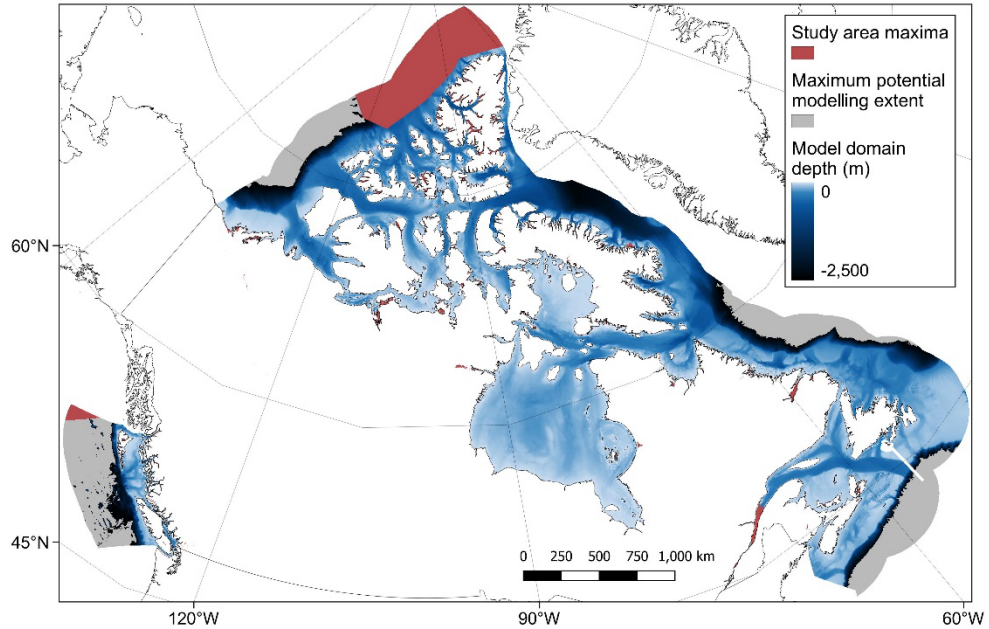
58   In the marine environment, organic carbon can originate from the fixation of carbon dioxide (CO_2)
59   by primary producers in the photic zone or via lateral transport from terrestrial sources (LaRowe
60   et al., 2020b). Organic carbon then passes through a variety of biotic and abiotic pathways being
61   consumed, transformed, respired or remineralised, with a large proportion converted back into
62   inorganic compounds, leaving only ~5% of marine production and less than 1% of earth's gross
63   production eventually reaching the seafloor (Middelburg, 2019; Hülse et al., 2017; Turner, 2015;
64   Bauer et al., 2013; Burdige, 2007). Once at the seafloor, a similarly complex process occurs on
65   and within the sediment, with a wide range of biotic, biochemical and physical processes all
66   influencing the rates of accumulation, remineralisation and resultant long term burial, with ~90%
67   of all carbon reaching the seafloor being remineralised (LaRowe et al., 2020b; Middelburg, 2018,
68   2019; Arndt et al., 2013). Even when considering this complex carbon cycle, the mass and
69   accumulation of organic carbon in surficial seabed sediments will still have a direct influence on
70   the scale of long-term carbon storage at the seafloor (LaRowe et al., 2020a; Middelburg, 2018).

71   Marine habitats are being increasingly recognised as contributors to nature-based climate change
72   mitigation (also known as nature-based climate solutions and natural climate solutions) due to
73   their ability to both fix CO_2 and store organic carbon for centennial to millennial timescales
74   (Macreadie et al., 2021; Hoegh-Guldberg et al., 2019). This "blue carbon" potential was initially
75   recognised in coastal vegetated habitats (i.e. mangrove, seagrass and saltmarsh) (Nellemann et
76   al., 2009; Duarte et al., 2005), but has more recently been applied to other habitats such as kelp
77   forests and unvegetated sediments (Luisetti et al., 2020; Raven, 2018; Avelar et al., 2017). There
78   is increasing evidence that human activities are influencing seabed sediment carbon stores from
79   both perturbations of upstream processes and physical impacts directly on the seafloor (Cavan
80   and Hill, 2022; Epstein et al., 2022; Keil, 2017; Bauer et al., 2013). For example, a recent study

81 estimated that the direct physical impacts from global fishing activities could cause considerable
82 remineralisation of seabed sediment organic carbon stocks back to CO_2 (Sala et al., 2021),
83 however the validity of the scale of these estimates has been called into question (Hiddink et al.,
84 2023; Hilborn and Kaiser, 2022; Epstein et al., 2022). By improving sediment carbon mapping
85 products, there may be opportunities to better research and design appropriate management
86 strategies to limit potential remineralisation from disturbance (Epstein and Roberts, 2022; Sala et
87 al., 2021; Luisetti et al., 2019).

88 Historically, studies measuring seabed sediment carbon stocks and accumulation rates had small
89 geographic scope, largely considering the ecological function, geological characteristics or
90 biochemical functioning at local to regional scales (see citations within LaRowe et al., 2020b;
91 Snelgrove et al., 2018; Middelburg, 2018; Burdige, 2007). In recent years, made possible by
92 modern machine learning and statistical spatial prediction techniques, there has been increasing
93 interest in estimating the size and distribution of carbon standing stocks and accumulation rates
94 at national to global scales to better understand natural carbon cycles and biological productivity,
95 and to identify the potential for improved management as a natural climate mitigation strategy
96 (Restreppo et al., 2021; Smeaton et al., 2021; Diesing et al., 2021; Atwood et al., 2020; LaRowe
97 et al., 2020b; Lee et al., 2019; Wilson et al., 2018; Avelar et al., 2017). Although global mapping
98 products can appear to give a complete understanding of seabed sediment organic carbon stocks
99 (Ludwig et al., 2023; Atwood et al., 2020; Lee et al., 2019), regional mapping studies which utilised
100 targeted data syntheses, refined spatial data products and finer resolution outputs, have shown
101 distinct spatial patterns in organic carbon distribution and disparate estimates of total standing
102 stocks when compared with these global studies. (Smeaton et al., 2021; Diesing et al., 2017,
103 2021; Luisetti et al., 2020; Wilson et al., 2018).

104 Canada has the world's longest coastline and approximately the seventh largest Exclusive
105 Economic Zone (EEZ) (Fig. 1), it could therefore be expected to contain a significant proportion
106 of the global stock of seabed sediment organic carbon. Data from recent global studies estimated
107 that the Canadian EEZ contains approximately 2.2 Gt of organic carbon in the top 5 cm and 48
108 Gt in the top meter of seabed sediments, equivalent to ~2.3% of total global marine sediment
109 carbon stocks covering around 1.3% of the area (Atwood et al., 2020; Lee et al., 2019). However,
110 these modelled estimates from global studies are at coarse spatial resolutions, have incomplete
111 coverage of the Canadian EEZ and contain very limited empirical data from within the Canadian
112 EEZ itself. Here, we conduct a systematic review of data on seabed sediment organic carbon
113 content across Canada and combine this with a synthesis and unification of best available data

114 on sediment composition, seafloor morphology, hydrology and chemistry in a machine learning
115 predictive mapping process, to construct the first high-resolution national assessment of
116 Canadian seabed sediment organic carbon stocks. To aid clarity, a workflow diagram of the
117 proceeding methods and results sections is shown in Figure 2.

118

119 **Figure 1. Map of the Canadian Exclusive Economic Zone (EEZ).** The study area spatial maxima (red) was defined
120 using best available bathymetry data and covers the entire sub-tidal portion of the Canadian EEZ (see high resolution
121 figure for further detail around the coastline and Section 2.1.1 for more details). This is overlayed by the maximum
122 potential modelling extent (grey) which indicates only those areas where data were present for all predictor variables
123 (see Section 2.1.7). Due to the distribution of available response data, the final modelling domain was limited to a depth
124 of 2,500 meters (see Section 2.4) and is indicated with the colour relative to the estimated depth, from 0 (light blue) to
125 -2,500 (black). Country outlines from World Bank Official Boundaries, available at
126 https://datacatalog.worldbank.org/search/dataset/0038272.

127

128

129

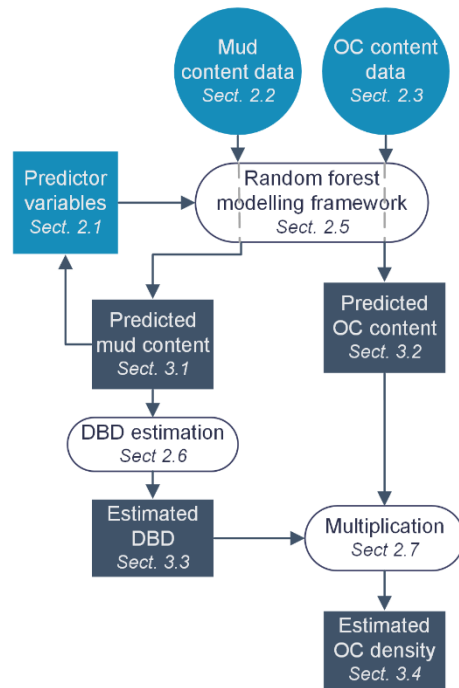



Figure 2. Study workflow diagram. Outline of the structure and linkages within the proceeding methods and results sections. Light blue shapes indicate input data; white ovals indicate data processes; dark shapes indicate output data; rectangles indicate raster data; circles indicate point data. OC = organic carbon; DBD = dry bulk density.

133

2. Methods

2.1 Predictor variables

2.1.1 Bathymetry

Best available contiguous Digital Elevation Model (DEM) data were combined and unified to a 200 m x 200 m equal area grid covering the Canadian EEZ (co-ordinate reference system (CRS) EPSG:3573 - WGS 84 - North Pole Lambert Azimuthal Equal Area Canada) (Table 1; see Appendix A2 for further details). Data were filtered to contain only sub-tidal areas (those cells with elevations of less than or equal to 0 m), with the resultant extent defined as the study area spatial maxima (Fig. 1).

143

144 **Table 1. Summary of predictor variables constructed for the Canadian EEZ.** For more information on methods
145 used to derive these layers see Section 2.1

Predictor variable	Unit	Region	Source	Native resolution	Temporal range
Bathymetry	m	BC	NRCan (2021)	10 m	NA
		Arctic	IBCAO V4.2 (Jakobsson et al., 2020)	200 m	NA
		Global	GEBCO (2022)	0.0042°	NA
Slope	°	Canada	This study	200 m	NA
Slope smoothed	°	Canada	This study	1 km	NA
Total curvature	rad/m	Canada	This study	200 m	NA
Total curvature smoothed	rad/m	Canada	This study	1 km	NA
BPI – fine	m	Canada	This study	200 m	NA
BPI – medium	m	Canada	This study	400 m	NA
BPI – broad	m	Canada	This study	400 m	NA
VRM – fine	-	Canada	This study	200 m	NA
VRM – medium	-	Canada	This study	200 m	NA
VRM – broad	-	Canada	This study	400 m	NA
Distance to shore	m	Canada	This study	200 m	NA
Bioregion	-	Canada	DFO (2022)	NA	NA
Distance to rivers – large	m	Canada	NRCan (2019)	1:15000000	NA
Distance to rivers – medium	m	Canada	NRCan (2019)	1:5000000	NA
Distance to rivers – small	m	Canada	NRCan (2019)	1:1000000	NA
Exposure proxy	-	Canada	This study	200 m	NA
SPM (surface)	g/m^3	Global	Copernicus (2022b)	4 km	2007 – 2019
Wave velocity (seafloor)	m/s	Arctic	Copernicus (2022a)	3 km	2007 – 2019
		Global	Copernicus (2022c)	0.2°	2007 – 2019
Mean current velocity (seafloor)	m/s	BC	Peña et al. (2019)	3 km	2007 – 2019
		Salish Sea	SalishSeaCast ERDDAP v19-05*	500 m	2007 – 2019
		Arctic & Atlantic	ANHA12 (Hu et al., 2019)[†]	0.0833°	2007 – 2019
Temperature (seafloor)	°C	BC	Peña et al. (2019)	3 km	2007 – 2019
	°C	Salish Sea	SalishSeaCast ERDDAP v19-05*	500 m	2007 – 2019
	°C	Arctic & Atlantic	ANHA12 (Hu et al., 2019)[†]	0.0833°	2007 – 2019
Salinity (seafloor)	ppt	BC	Peña et al. (2019)	3 km	2007 – 2019
		Salish Sea	SalishSeaCast ERDDAP v19-05*	500 m	2007 – 2019
		Arctic & Atlantic	ANHA12 (Hu et al., 2019)[†]	0.0833°	2007 – 2019
Ice thickness (surface)	m	Arctic & Atlantic	ANHA12 (Hu et al., 2019)[†]	0.0833°	2007 – 2019
Ice concentration (surface)	%	Arctic & Atlantic	ANHA12 (Hu et al., 2019)[†]	0.0833°	2007 – 2019
Dissolved oxygen (seafloor)	mol/m^3	Global	Bio-ORACLE V2.2 (Assis et al., 2018)	0.0833°	2000 – 2014
Primary production (surface)	g/m^3/d	Global	Bio-ORACLE V2.2 (Assis et al., 2018)	0.0833°	2000 – 2014
Chlorophyll concentration (surface)	mg/m^3	Global	Bio-ORACLE V2.2 (Assis et al., 2018)	0.0833°	2000 – 2014
Max current velocity (seafloor)	m/s	Global	Bio-ORACLE V2.2 (Assis et al., 2018)	0.0833°	2000 – 2014

146 Notes: BC = British Columbia; BPI = Benthic position index; VRM = Vector ruggedness measure; SPM = Suspended
147 particulate matter. *See https://salishsea.eos.ubc.ca/erddap/index.html; Soontiens and Allen (2017); Soontiens et al.
148 (2016). [†]See: https://canadian-nemo-ocean-modelling-forum-community-of-
149 practice.readthedocs.io/en/latest/Institutions/UofA/Configurations/ANHA12/index.html

150

151

152 ### 2.1.2 Benthic terrain features

153 A set of 10 benthic terrain features were constructed from the unified bathymetric layer (Table 1).
154 As benthic terrain measures use data on the depth of a location relative to the depth of
155 surrounding cells up to a given distance, bathymetric data within a given buffer outside the study
156 area maxima were included as needed to avoid edge effects in each terrain feature. Slope and
157 total curvature were calculated using the *terra.terrain* (Hijmans, 2022) and *spatialEco.curvature*
158 (Evans and Murphy, 2021) functions respectively. As these measures can be particularly sensitive
159 to artifacts from the DEM models and projections, they were constructed at two resolutions – the
160 native 200 m resolution, and after aggregating the bathymetry by 5-fold to 1 km x 1 km (termed
161 "smoothed"). Smoothed layers were disaggregated back to a 200 m resolution to maintain
162 uniformity across predictor layers.

163 Benthic position index (BPI) and vector ruggedness measures (VRM) were each calculated using
164 the *MultiscaleDTM* package at 3 different levels to capture both small local features and larger
165 spatial variation in terrain (Maxwell and Shobe, 2022; Ilich et al., 2021). Benthic position index
166 was calculated as the difference between the depth of a focal cell and the mean of cells contained
167 in annulus shaped window of 0.2 km to 5 km (BPI fine), 2 km to 25 km (BPI medium) and 4 km to
168 100 km (BPI broad). Vector ruggedness was measured by considering variation in the depth
169 surrounding each cell within square windows of width 1 km (VRM fine), 5.8 km (VRM medium)
170 and 11.6 km (VRM broad). Due to extremely inhibitive computational times when calculating VRM
171 broad, BPI medium and BPI broad at 200 m resolution, for these features the bathymetric layer
172 was first aggregated to a 400 m resolution before feature calculation, and then disaggregated
173 back to 200 m to maintain uniformity.

174

175 ### 2.1.3 Predictors describing the geographic setting

176 The geographic setting of each cell was described by its distance to shore and rivers, its broad
177 bioregional classification, and a proxy measure for exposure describing the degree of exposition
178 vs. shelteredness (Table 1). The geographic setting features are also influenced by the values of
179 surrounding pixels, therefore appropriate buffers were also applied to the processing of these
180 layers to avoid edge effects. Distance to shore was measured by the Euclidian distance to the
181 nearest land cell (indicated by an 'NA' value in the bathymetry layer), while bioregion was defined
182 by the Fisheries and Oceans Canada Federal Marine Bioregions classification (DFO, 2022). The

183   bioregion polygons were edited to include all bathymetry cells and re-classified with an integer
184   scale of 1 to 12 from east to west.

185   CanVec is a digital cartographic reference product produced by Natural Resources Canada
186   (NRCan) which includes the location of rivers across Canada at three mapped scales (NRCan,
187   2019). Firstly, the coarsest scale data (1:15,000,000) was projected onto the CRS of the
188   bathymetry layer and converted from polylines to a 2 km resolution raster. A 2 km buffer was
189   added around each river to ensure overlap of river mouths with the bathymetry data. The resultant
190   raster layer was resampled onto the bathymetry raster and the grid distance of each bathymetry
191   cell to the nearest river-mouth cell was calculated using the *terra.gridDist* function (Hijmans,
192   2022). This was then repeated for the medium scale (1:5,000,000) and fine scale (1:1,000,000)
193   layers with each river raster overlayed with the previous coarser scale layer to ensure all rivers
194   were included as the scales decreased.

195   To approximate the exposure setting of each cell, data on the mean distance from shore of
196   surrounding cells was used to construct a proxy value of fetch. Using the *terra.focal* function
197   (Hijmans, 2022), the mean distance to shore of surrounding pixels was calculated in square
198   windows of width 10 km, 20 km, 50 km, 100 km, 175 km and 250 km. Due to extremely inhibitive
199   computational times when calculating these values at the two largest distances, the distance to
200   shore layer was first aggregated to a 400 m resolution before focal calculations of these
201   components, and then disaggregated back to 200 m to maintain uniformity. The maximum value
202   in each layer was then set to the relative window size, and all data in each layer normalised
203   between 0 and 1. The mean of all layers was then calculated which resulted in continuous
204   measure of relative exposure/shelteredness ranging from 0 (highly sheltered) to 1 (highly
205   exposed).

206

207   2.1.4 Satellite derived predictors

208   Using data from the Copernicus Marine Data Store, two layers were created approximating the
209   mass of suspended particulate matter in surface waters and the orbital velocity of waves at the
210   seafloor. Data on suspended particulate matter (SPM) in surface waters across Canada from
211   2007 to 2019 was extracted in netCDF format from ACRI-ST (Sophia Antipolis, France)
212   company's global Bio-Geo-Chemical products at 4 km spatial resolution and a monthly temporal
213   resolution (Copernicus, 2022b). The climatological mean across this entire period was then
214   calculated for each cell and the netCDF converted to a raster for further processing. Due to the

215 complex nature of the Canadian coastline and the large disparity in spatial resolution of the
216 satellite data product (4 km) and the layers created above (200 m), the satellite raster layer was
217 allowed to extrapolate by 1 cell in its native resolution by taking the mean value of neighbouring
218 pixels. This allowed better overlap of satellite layers with the study area maxima at the coastline
219 but limited over-extrapolation. The raster layer was then reprojected to the equal area CRS and
220 resampled onto the bathymetry layer using cubic-spline interpolation. Due to a lack of consistent
221 SPM data recorded in the northern Arctic Basin, this portion of the data layer was manually
222 removed within QGIS.

223 To calculate the estimated orbital velocity of waves at the seafloor, two satellite wave data
224 products were combined with the unified bathymetry layer as constructed above. Hourly data from
225 2007 to 2019 on the significant wave height (H_s; VHM0) in meters, and primary wave swell mean
226 period (T_z; VTM01_SW1) in seconds, were extracted from the 0.2° resolution Global Ocean Wave
227 Reanalysis (WAVERYS) produced by Mercator Océan International (Copernicus, 2022c) and the
228 3 km resolution Arctic Ocean Wave Hindcast produced by MET Norway (Copernicus, 2022a). All
229 data were processed as the SPM data layer (except for lack of removal of the Arctic basin data),
230 and converted to an estimate of orbital wave velocity at the seafloor (U_{rms}; measured in m s^{-1})
231 using the following equation from Soulsby (2006);

232
$$U_{rms} = \left(\frac{H_s}{4}\right)\left(\frac{g}{d}\right)^{0.5} \exp\left\{-\left[\left(\frac{3.65}{T_z}\right)\left(\frac{d}{g}\right)^{0.5}\right]^{2.1}\right\} \tag{1}$$

233 where g is the acceleration due to gravity (9.806 m/s^2) and d is the water depth (m), taken as the
234 unified bathymetry layer multiplied by -1, and all values less than 1 meter depth rounded up to
235 the nearest meter (as needed for the above calculation). The resultant Arctic orbital velocity data
236 layer was then bias corrected to the global orbital velocity data layer utilising the *qmap* package
237 with quantile mapping using a smoothing spline (Gudmundsson et al., 2012). Finally, the two data
238 layers were overlayed with the regional Arctic data taking priority over the global data where
239 available.

240

241 2.1.5 Ocean circulation model predictors

242 To incorporate best available regional evidence, data on the mean surface ice cover, seafloor
243 salinity, temperature and current velocity was collated from three different ocean circulation model
244 products covering different regions of Canada (Table 1; see Appendix A3 for further details).
245 Three-dimensional data for salinity, temperature, u-velocity (eastward) and v-velocity (northward)

246   were extracted from each model and the climatological mean across all time points between 2007-
247   2019 was calculated. For each horizontal cell, data were only retained from the lowest vertical
248   cell within a given position (i.e. the cell which contacts the seafloor). Individual model outputs
249   were then converted to spatial point data using the cell centroid positions and transformed to the
250   unified CRS. Point data was then converted to rasters with the respective resolution of each
251   model, and the mean value taken if two points from the same model lay within a single raster cell
252   as an artifact of reprojection. As the Arctic-Atlantic model (ANHA12) has a varying horizontal
253   resolution, point data were rasterized using the smallest resolution of the original model (1.6 km)
254   and then interpolated using the *gstat* package (Gräler et al., 2016) and a nearest neighbour
255   interpolation method (including cells for land within the original model grid to supress
256   extrapolation). For all three models, mean current velocity was then calculated as the root mean
257   square of the u-velocity and v-velocity values in each cell. Finally, as carried out for the satellite
258   data layers, each raster was allowed to extrapolate by one cell in its native resolution (or for the
259   case of the ANHA12 model – its median resolution) and resampled onto the 200 m bathymetry
260   grid using cubic-spline interpolation. The three rasters were then combined with data only being
261   assigned to the spatial extent of the respective bioregions as defined in Section 2.1.3. Although
262   this means that different model products were used to measure the same predictor variable in
263   different regions, which can create biases, the bioregion predictor variable was included as a co-
264   variate in all models which included the ocean circulation variables, thus allowing for interactive
265   effects and accounting for differences in circulation model structures. Combining different models
266   can also create edge-effects, however the Arctic-Atlantic model is entirely spatially distinct so
267   contains no common edges with other models. The only significant edge between the remaining
268   two models lies at the mouth of the Juan de Fuca Strait and minimal disparity was seen (with the
269   other common edge occurring in the narrows of Johnson Strait).

270   Predictor layers describing the mean concentration and thickness of sea ice for the same temporal
271   period across the Arctic and Atlantic were also derived from the ANHA12 model. Processing of
272   model data and spatial rasters was conducted as above, except a value of zero ice concentration
273   and thickness was applied to all cells across the British Columbia Pacific bioregions.

274

275   2.1.6 Global model predictors

276   Four additional predictor variables were derived from Bio-ORACLE version 2.2 – a global unified
277   marine environmental data-layers collation which gives climatological mean values at 1/12th

278   degree resolution, for 2000-2014 and a wide-range of environmental variables (Assis et al., 2018).

279   Although these datasets are of lower resolution when compared to the regional data used above,

280   based on previous research there were some additional variables not available from the regional

281   circulation models which were considered potentially important for carbon modelling (Diesing et

282   al., 2021; Atwood et al., 2020). Three described the oceanographic chemistry/biology – namely

283   primary production and chlorophyll content of the surface water column, and dissolved oxygen

284   concentration at the seafloor. The fourth predictor was an additional measure of current velocity

285   (maximum current velocity), which was selected on top of the previously derived mean values

286   because current velocity has been identified as a particularly strong predictor within previous

287   seafloor sediment composition and carbon content predictive mapping studies (Gregr et al., 2021;

288   Diesing et al., 2021; Mitchell et al., 2019). Raster data were downloaded from the Bio-ORACLE

289   website and processed as the satellite data layers (i.e. allowed to extrapolate by 1 cell in its native

290   resolution by taking the mean value of neighbouring pixels, reprojected to the unified equal area

291   CRS and resampled to the unified 200 m grid using cubic-spline interpolation).

292

293   2.1.7 Final collation of predictor variables

294   The resulting 28 predictor variable raster layers were combined into a single raster stack and any

295   cells containing NA values removed, leaving only those cells which contained values across all

296   predictor layers. The remaining cells covered 92.3% of the subtidal zone of the Canadian EEZ

297   and delineated the maximum potential modelling area (Fig. 1). The final predictor variable layers

298   are shown in the Supplement.

299

300   **2.2 Sediment mud content data**

301   Empirical point data on seabed sediment mud content across the Canadian EEZ were extracted

302   from two sources (Enkin, 2023; NRCan, 2022) (see Appendix A4 for further information). Data

303   were only retained if they originated from within the top 30 cm of the sediment and had associated

304   geographic position information (latitude-longitude co-ordinates; lat-lon). Data were further filtered

305   by excluding those where the sum of mud, sand and gravel content was greater than 102% and

306   lower than 98% - to allow for rounding errors but to exclude invalid data. Data were also excluded

307   if samples/sub-samples were not present from at least the top 1 cm to 5 cm below the sediment

308   surface within a given sampling event. After data filtering, the mean percentage of mud was taken

309   across replicates/sub-samples, leaving a single value for each sampling event. We chose to
310   concentrate on sediment mud content as this has previously been identified as the key sediment
311   composition component from a number of related carbon mapping studies (Smeaton et al., 2021;
312   Diesing et al., 2017, 2021; Pace et al., 2021; Wilson et al., 2018). Finally, mud content data were
313   projected onto the CRS of the predictor layers and only retained where overlap occurred. This led
314   to a final dataset of 19,730 samples (Fig. B1).

315

316   **2.3 Organic carbon content data**

317   2.3.1 Organic carbon data collation and extraction

318   Data on the percent organic carbon content within dried surface sediments (%OC) was collected
319   from three different structured searches. Firstly, a systematic literature review was conducted
320   through Web of Science and Scopus. Both searches were conducted on the 21st September 2022.
321   Within Web of Science, its "Core collection" was searched via the field "Topic", which examines
322   a paper's title, abstract, author, keywords and "keywords plus". Within Scopus, the search was
323   run via the field "Title-Abs-Key", which scans a paper's title, abstract and keywords. Within both
324   databases the same search string was used:

325   ("organic carbon" OR "organic matter" OR "organic content" OR TOC OR TOM) AND (coast* OR
326   sea* OR ocean* OR estuar* OR marine OR gulf) AND (sediment* OR mud* OR sand* OR clay*
327   OR silt* OR gravel* OR seabed) AND Canad*

328   All articles identified from the searches were exported into a single Zotero library and duplicates
329   removed, leaving 1,581 results. Screening was conducted via a hierarchical process that first
330   assessed the title, then abstract and finally full text. At each stage an article was assessed against
331   the inclusion criteria described below, with those considered relevant or of unclear relevance
332   passing to the next level of assessment.

333   The inclusion criteria were defined as: 1) Study conducted on subtidal seabed sediments (those
334   concerning rock, shale or fauna were not included); 2) Physical samples collected using a seabed
335   sediment sampling device (e.g. cores or grabs – sediment-trap samples were not included); 3)
336   Samples from within the Canadian EEZ; 4) Studies concerning the chemical composition of the
337   sediment; 5) Organic carbon content (%) directly measured after separation of organic and
338   inorganic components (e.g. by acidification). After the title screening stage 242 articles remained,
339   followed by 123 remaining after abstract screening, and a final set of 49 articles left for data

340 extraction after review of the full text. Four additional primary literature papers were added based

341 on expert advice. This included two large data collation studies, one concentrating on the Arctic

342 (CASCADE; Martens et al., 2021) and one having global scope (MOSAIC v2; Paradis et al., 2023)

343 The second structured search was conducted on the Canadian Federal Science Libraries Network

344 – a repository which contains departmental publications, reports and data sets from seven

345 science-based Canadian government departments. The search was carried out on the 7th

346 November 2022 using the same search string as for the primary literature and querying all fields.

347 The search led to only 178 results and therefore each result was assessed individually against

348 the selection criteria first by their abstract and then by a full text assessment, leading to data

349 extraction from 15 reports. The third search was carried out on the 15th November 2022 using

350 GEOSCAN – the NRCan bibliographic database for scientific publications. As GEOSCAN does

351 not allow search strings containing "AND", the search was conducted on all fields using only the

352 terms: "organic carbon" OR "TOC" OR "OC"; leading to 655 search results. The metadata of all

353 entries was exported as a text file and further refined using a secondary manual search for the

354 remainder of the search terms listed above within Microsoft Excel. This led to a final set of 233

355 results, 178 which were excluded by screening of the title, and a further 51 excluded by abstract

356 or full text screening, leaving 4 reports for data extraction.

357 In total, these three structured searches of primary literature and government reports led to 72

358 publications for data extraction. As well as data on the %OC, metadata extracted included the

359 maximum depth of sample into the sediment (cm), geographic position (lat-lon), sample ID, year

360 of sampling (approximated as publication year where not clearly stated), sampling method (e.g.

361 multicorer, Van Veen grab) and water depth of sample site (where recorded). Data were extracted

362 from data tables or supplementary databases when available, otherwise the PlotDigitizer online

363 application was used to extract data from graphical products. Where possible data were extracted

364 on the %OC in different depth-layer sub-samples through a single core-sample up to 50 cm,

365 otherwise a single mean value was taken.

366 Additional to data collated through the structured searches, %OC data were also extracted from

367 PANGAEA – a global data repository for geographic earth-system data (PANGAEA®, 2022). A

368 data search across all topics was conducted on the 25th October 2022 using the same search

369 terms as for the structured search, except for removal of the term "*Canad**". The geographic

370 extent of the results was instead delineated using the spatial tool within PANGAEA which allows

371 results to be filtered by the geographic co-ordinates of a square/rectangular extent. Overall, this

372 led to a total of 1,489 potential datasets. All relevant data within these datasets were exported

373   using the Data Warehouse Download tool within Pangaea. Based on expert knowledge, two
374   additional PANGAEA datasets were added to the output from published global %OC data-
375   syntheses (Atwood et al., 2020; Seiter et al., 2004). Lastly, where the date of the sample was not
376   recorded, the sampling year was manually added by further exploring the metadata or cited
377   studies. To align the PANGAEA data with the systematic review data, PANGAEA data points
378   were excluded if: 1) they lacked data on %OC; 2) they lacked metadata on the depth of a sample
379   within the sediment; 3) if the sample originated from greater than 50 cm below the sediment
380   surface; or 4) metadata on the elevation/water depth indicated sampling above the subtidal.
381   Additionally, metadata within PANGAEA were coalesced where necessary (due to different
382   names being given to the same data type), and mean values of %OC taken if replicates were
383   measured within a single sub-sample.

384   All organic carbon data were converted into spatial point data, transformed to the unified equal
385   area CRS and masked by the predictor variable's maximum model area to leave only overlapping
386   data. Additionally, values were only retained from the sampling year 1959 and onwards. The extra
387   year was included when compared to the mud content data because there were some widescale
388   surveys undertaken across the Labrador Sea in 1959 which was lacking from any additional %OC
389   datasets. While this large temporal extent may add uncertainty in relation to the quality and
390   uniformity of the response data, similar extents have been used by previous global mapping
391   studies (Atwood et al., 2020; Lee et al., 2019; Seiter et al., 2004) and, 72% of the %OC data
392   within this study were sampled after 1980 and 55% after 2000. The larger temporal extent also
393   allows for the inclusion of a larger frequency and wider spatial extent of data, therefore potentially
394   improving robustness of our spatial predictive models. In total our %OC dataset contained 2,518
395   point-samples (Fig. B2) and 3,308 sub-samples across different depth layers within cores.

396

397   2.3.2 Organic carbon data processing

398   Due to commonly adopted uneven sampling distributions within single core samples (i.e. more
399   sub-samples towards the top of the core), where sub-sample data were present on the %OC in
400   different depth-layers these were converted into weighted cumulative means assuming linear
401   distribution between sub-samples. Additionally, there was large variation in the maximum
402   sediment depth of point-samples, ranging from %OC measures from only the top 1 cm of
403   sediment, to values up to the chosen data extraction limit of 50 cm deep. We chose to standardise
404   all samples to 30 cm depth as only 6% of the point-samples covered sediment depths below this

layer and because 30 cm is a commonly suggested carbon stock accounting depth for terrestrial soil and marine sediment habitats in both carbon accrediting methodologies and greenhouse gas inventories (VERRA, 2020; IPPC, 2019).

To estimate the cumulative mean of %OC at 30 cm for all individual point-samples, we created a transfer function using a generalised additive mixed model (GAMM) smoothing spline. It is generally expected that the %OC in marine sediments decreases with depth within the seafloor (Middelburg, 2018); we used the collated data above to approximate a mean trend for this study. Firstly, only those data that contained at least five sub-sampled depth layers were retained for modelling as fitting distributions to those with fewer points would likely be invalid. This left 183 unique samples with 2,640 weighted cumulative mean sub-samples for model construction. Cumulative mean %OC data were arcsin transformed (arcsin{$\sqrt{}$ [%OC/100]}; a commonly adopted transformation for percentage data), and a simple GAMM model applied with sub-sample sediment depth as the fixed factor modelled with a cubic regression spline and sample ID as the random factor. The GAMM model was fitted using the *mgcv* package; a scaled-t distribution family was used for heavy tailed Gaussian-like data, the number of basis dimensions was set to 20 and smoothing parameter estimation was conducted by Restricted Maximum Likelihood (REML) (Wood et al., 2016). Model validation was carried out using visual assessment of diagnostic plots of residuals, as well as observed vs fitted values. Significance of the sampling depth smoothing spline was assessed by an analysis of variance (ANOVA) with a chi-squared test comparing the full GAMM model to a null GAMM model containing only the random factor and the intercept (see Appendix C for results). The difference between estimated deviance explained in the full and null models was also used to approximate the variance explained by the fixed and random factors. To create a transfer function, the cumulative %OC was predicted from the mean fixed effects of the GAMM model at sediment depths from 0 – 30 cm at 0.1 cm intervals. The predictions were then back-transformed to percentage data and the cumulative mean %OC at each depth was converted to an inverse proportion of the mean across 30 cm. Overall, this gave an estimated proportional conversion factor from the cumulative mean at any given depth to an expected mean across 30 cm (Appendix C).

All point-sample data from PANGAEA and the systematic review were combined, corrected to weighted cumulative means where sub-samples were present, checked for duplication, and unified to a mean %OC value of the top 30 cm of sediment using the above transfer function. One outlier was removed from the dataset as it was reported to have a carbon content twice that of any other sample within the dataset. Finally, for further analyses %OC data were arcsin

438   transformed due to a highly right skewed distribution and its application within similar modelling
439   exercises (Smeaton et al., 2021; Diesing et al., 2017).

440

441   **2.4 Final model domain selection**

442   After visual assessment of the coverage of both the mud content and %OC data, the final model
443   domain was limited to a water depth of 2,500 meters. This depth limit (as delineated by the
444   bathymetry predictor layer) encompassed 99.95% of mud content point data (Fig. B1) and 99.3%
445   of %OC data (Fig. B2). The predictor layer raster stack was filtered with all cells deeper than
446   2,500 meters excluded from the model domain. This final model domain covers 4,489,235 km^2
447   which is 78.4% of the EEZ or 92.6% of the seafloor area above 2,500 m (Fig. 1).

448

449   **2.5 Random forest modelling**

450   For predictive mapping we adopted random forest machine learning techniques due to their
451   flexibility regarding violations of traditional statistical assumptions, ability to handle a range of data
452   types and predictor variables and elucidate both drivers of model response and predictions of
453   uncertainty, as well as their successful application in previous similar modelling tasks (Diesing et
454   al., 2017, 2021; Pace et al., 2021; Atwood et al., 2020; Wilson et al., 2018). Mud content and
455   organic carbon content (%OC) were both modelled using the following framework. Firstly, each
456   response variable was overlain onto the predictor variable grid and the mean values were taken
457   if more than one data-point fell within a single raster cell. All predictor variable data were then
458   extracted for each response dataset; however, the three biological/biochemical predictor
459   variables (primary production, chlorophyll concentration and dissolved oxygen) were only used
460   within the %OC model as they are not expected to drive variation in physical sediment properties
461   (Restreppo et al., 2021; Gregr et al., 2021; Graw et al., 2021; Mitchell et al., 2019).

462   Contemporary research in spatial machine learning techniques have highlighted that robust
463   spatial cross-validation (CV) strategies and predictor variable selection processes are essential
464   to calculate valid performance metrics, limit overfitting and construct reliable spatial predictions
465   (Zhang et al., 2023; Ludwig et al., 2023; Meyer and Pebesma, 2022; Meyer et al., 2019). Details
466   of the methods used to ensure appropriate cross-validation design and feature selection are
467   discussed in Appendix A5. For both response variables, following identification of an appropriate
468   35-fold spatial CV structure, a single fold was held-back as testing data, with all other data

469   retained for model training. To ensure an absence of duplication between the training and testing

470   data, 34 spatial CV folds were reconstructed on the training data (i.e. all training data assigned to

471   one of 34 validation sets). Using these CV folds, the *CAST.ffs* function (Meyer et al., 2023) was

472   then used to run a forward predictor variable selection processes with appropriate spatial

473   considerations (see Appendix A5 for further information).

474   Following variable selection, hyperparameter tuning was conducted on the hyperparameters *mtry*

475   (the number of variables to randomly sample as candidates at each split) and *min_n* (the number

476   of observations needed to keep splitting nodes); with the *trees* hyperparameter (the number of

477   random forest trees to construct and take mean predictions across) set to 1,000 (Probst et al.,

478   2019). 11 potential combinations of hyperparameters were selected using a semi-random Latin

479   hypercube grid (Kuhn and Silge, 2023; Kuhn and Wickham, 2020). The tuning process fitted

480   separate models across all CV folds and hyperparameter combinations (i.e. 34 CV folds x 11

481   hyperparameter options = a total of 374 models) (Kuhn and Silge, 2023; Kuhn and Wickham,

482   2020). The performance of each of the 11 hyperparameter combinations was assessed by

483   calculating the root mean squared error (RMSE) on predictions of the validation data across all

484   CV folds, with the optimal hyperparameter combination selected as that with the lowest RMSE

485   (Meyer et al., 2019, 2023). After selection of the best performing hyperparameter combination, a

486   single last-fit model was constructed on the entire training set and evaluated on the held-back test

487   set (Kuhn and Silge, 2023; Kuhn and Wickham, 2020), with the absence of overfitting determined

488   by the RMSE and R^2 of the last-fit model falling within the range of those found across CV folds

489   with optimal hyperparameters. Final model performance metrics (RMSE and R^2) were then

490   calculated using all predictions of validation data from CV folds (with optimal hyperparameters)

491   and from predictions of the testing data from the last-fit model (Meyer et al., 2019, 2023).

492   Predicted values were then calculated across the entire model domain using the last-fit model

493   and the predictor variable raster stack (Kuhn and Silge, 2023; Kuhn and Wickham, 2020), and

494   cell-specific estimation of uncertainty was calculated using standard error on out-of-bag

495   predictions using infinitesimal jack-knife for bagging (Roy and Larocque, 2020; Wager et al.,

496   2014). Due to computational restraints when calculating predictions across the entire model

497   domain (which contains 112,230,871 cells), the predictor variable raster stack was split into 150

498   non-overlapping partitions by random sampling and both prediction and standard error estimates

499   made serially on each partition. All predictions were then merged to create a raster layer covering

500   the entire model domain (although edge effects were not expected between partitions, random

501   sampling without replacement across the entire domain was chosen to ensure its absence).

A cell-specific approximation of the upper and lower bounds of the 95% confidence interval (CI) was calculated by adding/subtracting the cell-specific standard error estimates, each multiplied by 1.96, from the mean predictions and then back transformed where needed (Kuhn and Wickham, 2020; Wager et al., 2014). After calculation, CI values were corrected where necessary – being bounded by 0 and 100. The resulting three raster layers from the mud content model were also used as available additional predictor variables when constructing the random forest models for %OC (Fig. 2). Although this gives the potential for data leakage if mud content and %OC data were from the same samples, we found only 31 occurrences (1.3% of OC samples) where direct spatial overlap occurred, and therefore do not consider that significant data leakage is present and no impact on variable importance or model performance calculations will be seen. Finally, a measure of relative predictor variable importance was calculated by fitting an additional single random forest model on all training data using optimal hyperparameters, and the predictor importance calculated on out-of-bag data through permutation of predictor variable values (for further details see Kuhn and Silge, 2023; Wright et al., 2016). Accumulated local effects (ALE) plots for the last-fit model were produced for the six predictor variables with highest importance in each model using the *iml* package (Molnar et al., 2018) to give a visual representation of the average effect of predictors on model prediction outcomes.

2.6 Estimating sediment dry bulk density

An estimate for the dry bulk density of the sediment (ρD – the mass of dried sediment per unit volume within the seafloor; g cm^{-3}) was constructed across the model domain based on the predictions of mud content from the random forest model (Fig. 2). We identified three published functions which describe the relationship between mud content and porosity (Φ; the proportion of sediment volume which is water) in seabed sediments. The following equations are respectively from Jenkins (2005), Diesing et al. (2017) and Pace et al. (2021):

$$\Phi = 0.3805 . mud + 0.42071 \tag{2}$$

$$\Phi = 0.4013 . mud + 0.4265 \tag{3}$$

$$\Phi = 10^{\{0.138 . \log_{10}(mud) - 0.486\}} \tag{4}$$

Due to each of these equations being approximations of the relationship between mud content and Φ, we chose to take the mean response. In all equations *mud* represents the predicted mean mud content values as calculated above, each expressed as a decimal proportion. For Equation

533   4, mud content was rounded up to the nearest 0.01 as lower values give unrealistic porosity

534   estimates. Sediment porosity can then be converted to an estimate of dry bulk density using the

535   following equation:

536     $\rho D = \rho S(1 - \Phi)$ (5)

537   where ρS is the grain density of seabed sediments in g cm^{-3}, which was set at the frequently used

538   constant approximation of 2.65 (Diesing et al., 2017, 2021; e.g. Pace et al., 2021; Lee et al., 2019;

539   Wilson et al., 2018; Kuzyk et al., 2017). Although this standard approximation of grain density is

540   not ideal, the variation under different environmental settings is generally found to be small when

541   compared to differences in %OC and porosity, therefore the values of grain density are not

542   expected to strongly drive variation in organic carbon density (Atwood et al., 2020; Lee et al.,

543   2019; Middelburg, 2019; Martin et al., 2015; Berner, 1982).

544   To incorporate uncertainty from our mud content predictions, estimates of dry bulk density were

545   also calculated from the cell-specific predictions of the lower and upper bounds of the 95% CI of

546   mud content. We used these derived lower and upper bounds of dry bulk density estimates as

547   best available approximations of uncertainty around the dry bulk density mean estimate values.

548   Equivalent approaches to estimating uncertainty have been used in other seabed sediment

549   carbon mapping studies (e.g. Diesing et al., 2017, 2023; Lee et al., 2019).

550

2.7 Estimating organic carbon standing stock

552   The organic carbon density (g cm^{-3}) is calculated by multiplying the %OC (expressed as a decimal

553   proportion) by the sediment dry bulk density (Fig. 2). For the final calculations, the respective

554   means, upper and lower uncertainty bounds were multiplied together to incorporate uncertainty

555   from both components. These compound uncertainties were used as best available

556   approximations of the lower and upper bounds of uncertainty around the estimates of mean

557   organic carbon density (akin to Diesing et al., 2017, 2023; Lee et al., 2019). To create a more

558   meaningful response value, organic carbon density was converted to kg m^{-3} (multiplied by 1000).

559   Finally, the organic carbon stock in each mapped cell can be calculated by multiplying the organic

560   carbon density by the reference sediment depth of this study (0.3 m) and the cell area (40,000

561   m^2) and converted to metric tonnes (divided by 1000). Overall, this allows estimates to be

562   calculated for the total values of organic carbon stock across different parts of model domain.

563

564

2.8 Rock substrate distribution case studies

The method followed in this study is similar to that used for many similar seabed sediment predictive mapping exercises in that it uses data only from sediment grab and core samples to build predictive maps across the model domain (Restreppo et al., 2021; Graw et al., 2021; Diesing et al., 2017, 2021; LaRowe et al., 2020a; Atwood et al., 2020; Lee et al., 2019; Mitchell et al., 2019; Wilson et al., 2018; Stephens and Diesing, 2015). One major limitation with this modelling approach is that areas of bedrock, which would have zero values for all sediment response variables, will not be recorded in these datasets. Therefore, the under representation of zero values in the response data could lead to an overestimate of organic carbon standing stocks as zero values are unlikely to be predicted from model outputs.

In the context of this study, information regarding the distribution of bedrock is lacking for many regions. We therefore use two regional case studies from the Pacific British Columbian EEZ and the Atlantic Scotian shelf and slope where recent publications have made estimated maps on the distribution of rock substrates (Philibert et al., 2022; Gregr et al., 2021). Each of these products was overlayed onto the final spatial predictions of sediment carbon densities and all cells set to zero where rock substrates were predicted. The proportional effect on the mean, upper and lower estimates of carbon stock was then calculated in each bioregion.

582

3. Results

3.1 Mud content predictive mapping

Of the 25 predictor variables available for mud content random forest modelling, 13 were selected in the optimal model (Fig. 3). Mean orbital velocity of waves at the seafloor and the mass of suspended particulate matter at the surface were the variables with highest importance (Fig. 3). Other variables with relatively high importance for predicting mud content included the exposure setting, ice thickness, distance to rivers, bathymetry, and benthic position indices (Fig. 3). Higher mud content was generally predicted in areas of low wave velocity, low exposure and close to but not directly adjacent to river mouths; with the effect of SPM and ice thickness less distinct, likely due to more complex interactive effects (Fig. 4).

593

594

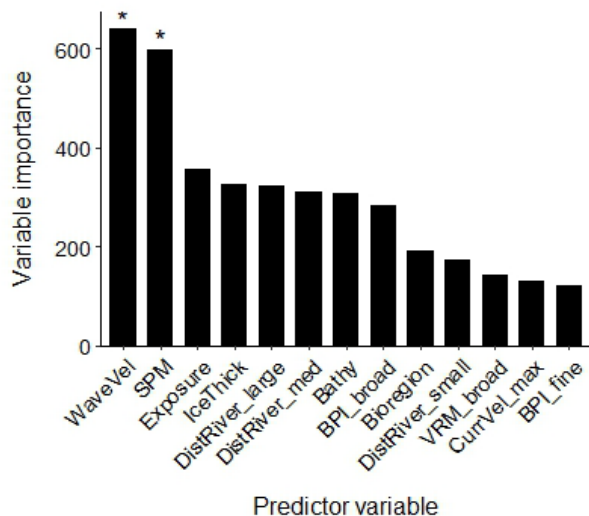



Figure 3. Predictor variable importance from random forest models of mud content in marine subtidal sediments. The y-axis is a unitless relative variable importance score for each model. Asterisks indicate the two initial predictors which were selected based on variable importance, with all other predictor variables selected using a forward selection process (see Appendix A5 for further details). WaveVel = Orbital wave velocity at the seafloor, SPM = Suspended particulate matter within the water column, BPI = Benthic position index, DistRiver = Distance to nearest river, IceThick = Sea ice thickness, Bathy = Bathymetry, VRM = Vector ruggedness measure, CurrVel = Current velocity at the seafloor.

602

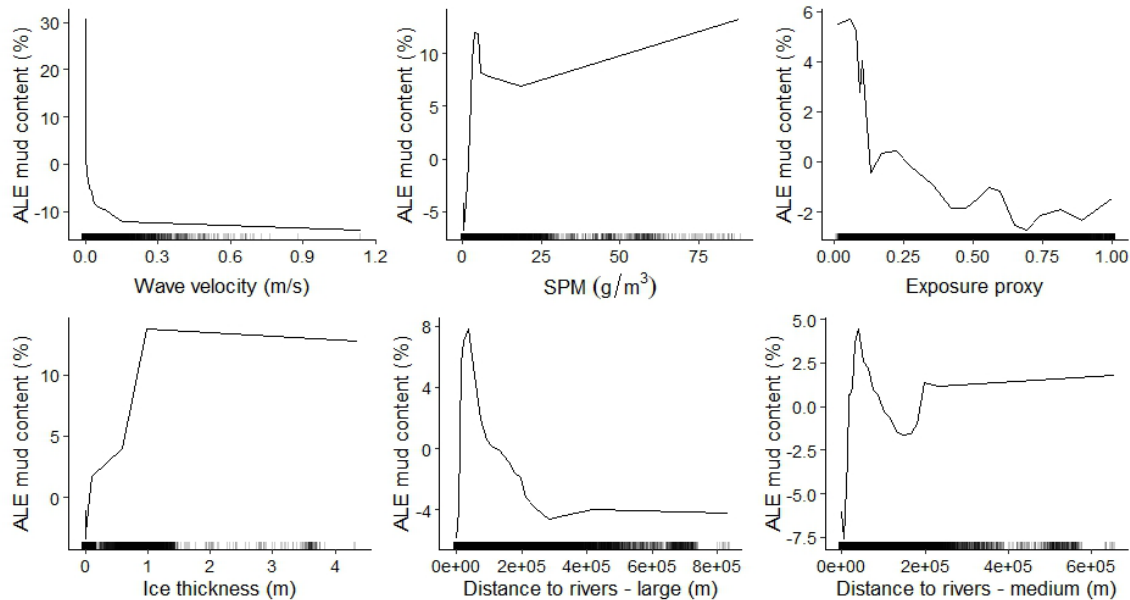



Figure 4. Accumulated local effects (ALE) plots for the six predictor variables with highest importance in the mud content random forest model. ALE (distributions dawn by lines) gives a visual representation of the average effect of the predictor variable on the response but does not indicate the influence of multi-way interactions which are inherent in random forest models. Rug plots (dashed marks at bottom) indicate the distribution of each variable within the training dataset. SPM = suspended particulate matter.

608

Areas with sediments dominated by mud (>75%) were predicted across the basins of many of the Pacific fjords, inlets and estuaries, and within the southern Salish Sea (Fig. 5). In the Arctic, mud dominated areas included large parts of the Canadian western Arctic as well as Hudson Bay. In the Atlantic, the Laurentian channel and deeper parts of the Scotian Shelf contained particularly high mud fractions (Fig. 5). Across the model domain, sediment in deeper areas on the continental slope was also highly dominated by mud (Fig. 5) Using robust spatial cross validation, the model was estimated to have an RMSE of 24.4% and R^2 of 0.60. The cell specific upper and lower 95% CI bounds are shown in Figure E1. On average the upper CI bounds were 28% larger than the mean and the lower CI bounds 20% less.

618

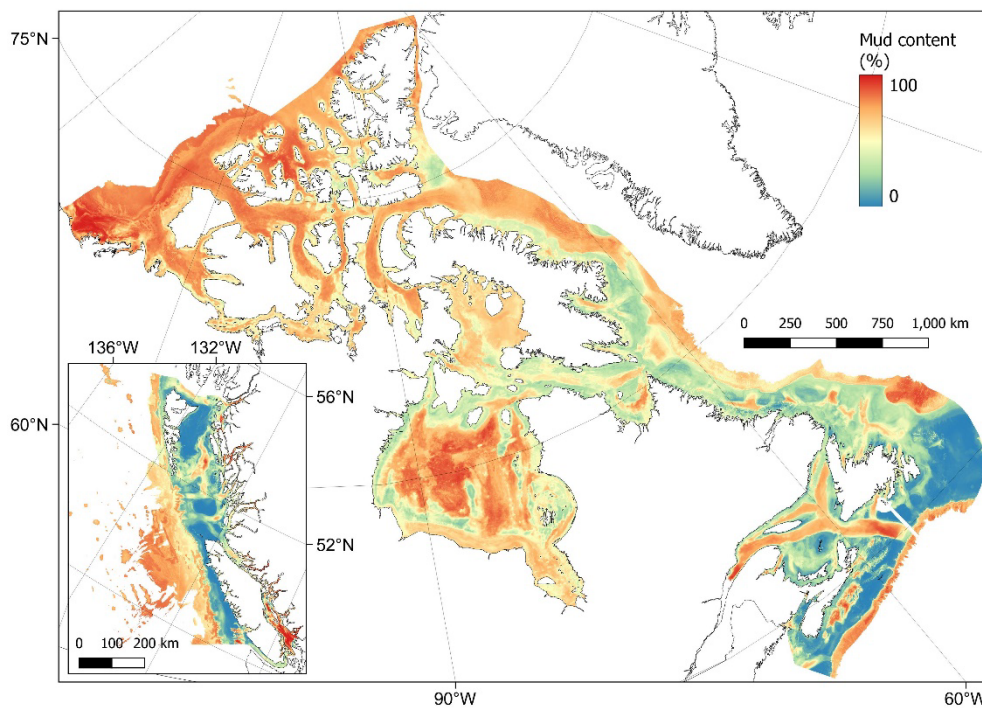



Figure 5. Predictive mapping of mud content (%) in subtidal marine sediments across the Canadian continental margin. The main plot shows the Arctic and Atlantic regions with the Pacific region inset. The 95% confidence interval bounds around the predicted means are shown in Figure E1. Labels indicating the locations of different areas mentioned within the text are shown in Figure B3. Country outlines from World Bank Official Boundaries, available at https://datacatalog.worldbank.org/search/dataset/0038272.

624

625

3.2 Organic carbon content predictive mapping

Eleven predictor variables were selected in the optimal organic carbon content (%OC) model (Fig. 6). The variables with highest importance in predicting %OC were the mud content layers constructed above (specifically the mean and lower bound), with all other predictors having less than half the relative importance of the mean mud predictions (Fig. 6). On average organic carbon content increased with predicted mud content and was generally higher in areas with low SPM concentrations, low exposure settings, close to but not directly adjacent to rivers, and at high water temperatures (Fig. 7).

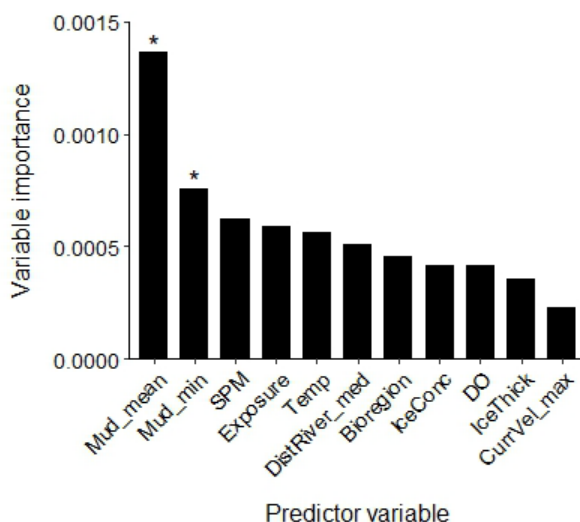



Figure 6. Predictor variable importance from random forest models for the organic carbon content in marine subtidal sediments. The y-axis is a unitless relative variable importance score**.** Asterisks indicate the two initial predictors which were selected based on variable importance, with all other predictor variables selected using a forward selection process (see Appendix A5 for further details).. Mud_min = Lower bound of 95% CI for mud content, SPM = Suspended particulate matter within the water column, Temp = Temperature, DistRiver = Distance to nearest river, IceConc = Sea ice concentration, DO = Dissolved oxygen at the seafllor, IceThick = Sea ice thickness, CurrVel = Current velocity at the seafloor.

644

645

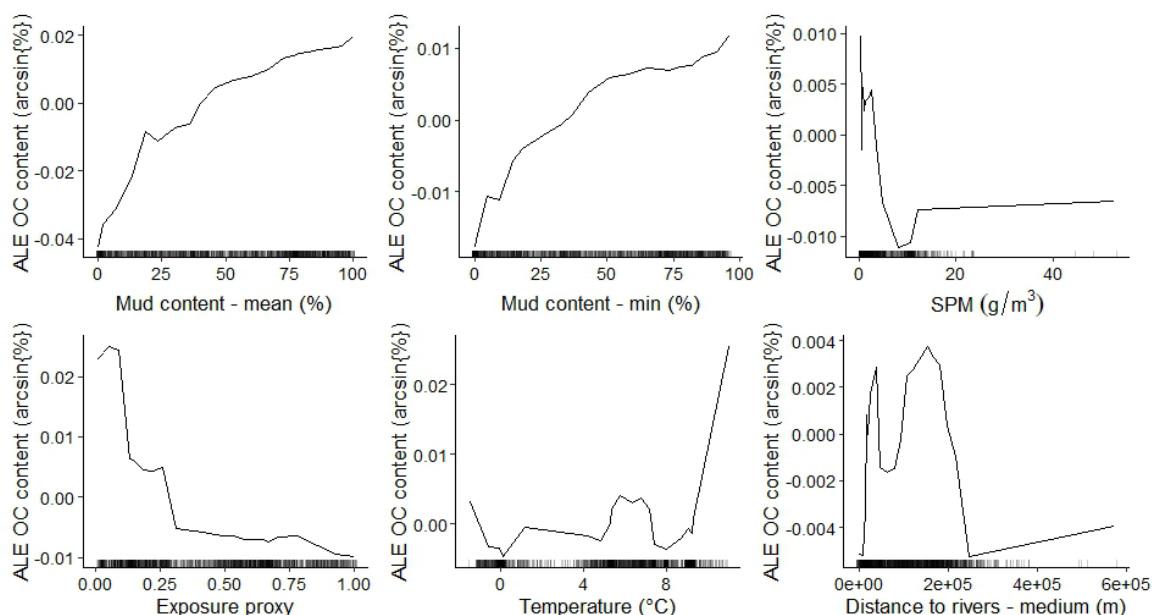



Figure 7. Accumulated local effects (ALE) plots for the six predictor variables with highest importance in the organic carbon (OC) content random forest model. ALE (distributions dawn by lines) gives a visual representation of the average effect of the predictor variable on the response but does not indicate the influence of multi-way interactions which are inherent in random forest models. Rug plots (dashed marks at bottom) indicate the distribution of each variable within the training dataset. SPM = suspended particulate matter.

The predictions of %OC ranged from 3×10^{-5} to 5.6% with an overall mean of 0.8 ± 0.3% (± SD). Areas with highest predicted %OC (>3%) were restricted to parts of the Pacific west coast fjords and channels, and in small parts of the inlets and bays on the east coast of Nova Scotia and around Passamaquoddy Bay in the Bay of Fundy (Fig. 8). High concentrations (i.e. >1%) were more widespread across these areas as well as covering much of the Beaufort Sea, western Baffin Bay and Foxe Basin in the Arctic, southern and central Hudson Bay, the Laurentian channel, coastal north Newfoundland and the central Scotian shelf in the Atlantic, as well as across the Salish sea and deeper areas to the south of the British Colombian Pacific continental margin (Fig. 8). Lowest %OC was predicted across shallower parts of the central Pacific shelf and near coast areas west of Vancouver Island (Fig. 8). Cross validation estimated an R^2 for the model of 0.58 and an RMSE of 0.09 arcsin{%OC}. Cell specific upper and lower 95% CI bounds are shown in Figure E2. On average the upper CI bounds were 42% larger than the mean prediction, and the lower CI bounds 33% less than the mean prediction.

664

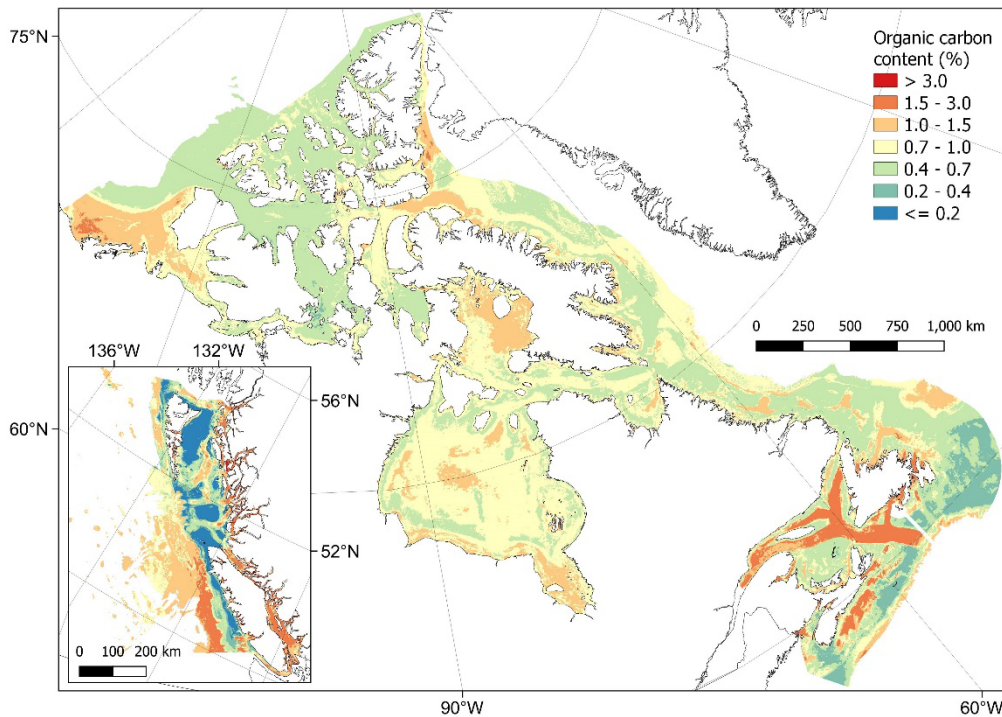



Figure 8. Predictive mapping of organic carbon content (%) in subtidal marine sediments across the Canadian continental margin. The main plot shows the Arctic and Atlantic regions with the Pacific region inset. The continuous variable is shown displayed in discrete colour bands to improve visualisation of highly right skewed data. The 95% confidence interval bounds around the predicted means are shown in Figure E2. Labels indicating the locations of different areas mentioned within the text are shown in Figure B3. Country outlines from World Bank Official Boundaries, available at https://datacatalog.worldbank.org/search/dataset/0038272.

3.3 Dry bulk density estimation

The dry bulk density of sediments was estimated using the predicted values of mud content from our random forest model, and previously published functions for conversions to porosity and dry bulk density (Fig. 2). Estimated values ranged from $0.67 – 1.61$ g cm^{-3} with a mean of 1.04 ± 0.21 g cm^{-3} ($\pm$ SD). As expected by its derivation, the spatial distribution of dry bulk density values was very similar to the mud content values predicted above (Fig. 5), i.e. lowest dry bulk density was estimated in mud dominated areas (Fig. 9). Cell specific upper and lower uncertainty bounds are shown in Figure E3. On average these bounds were 6.2% larger and 6.0% lower than the cell-specific mean estimate.

681

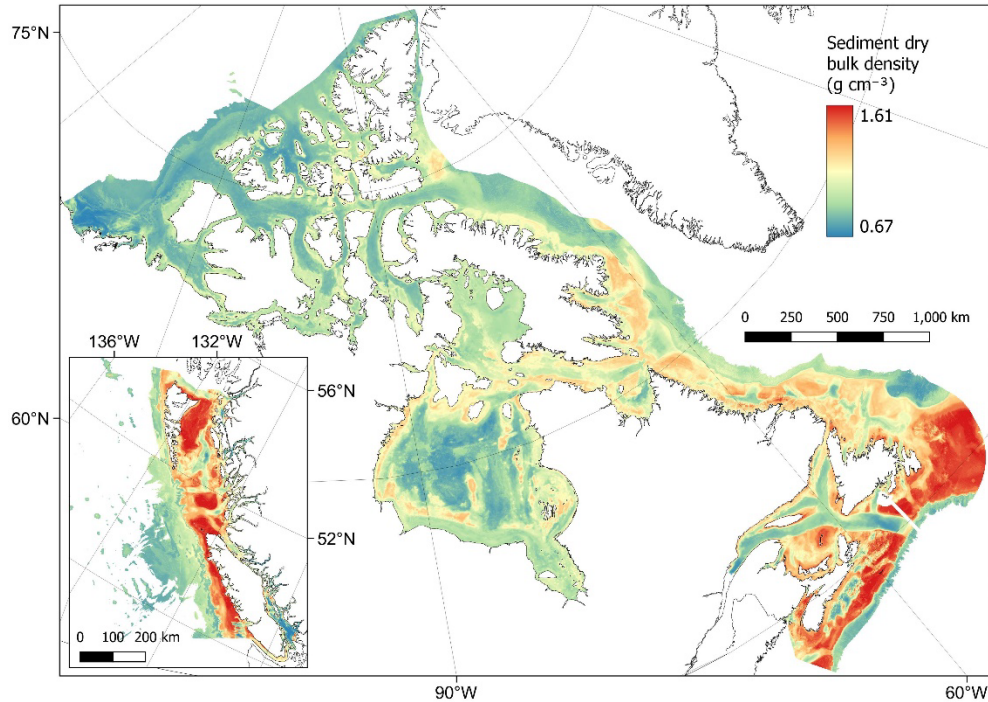



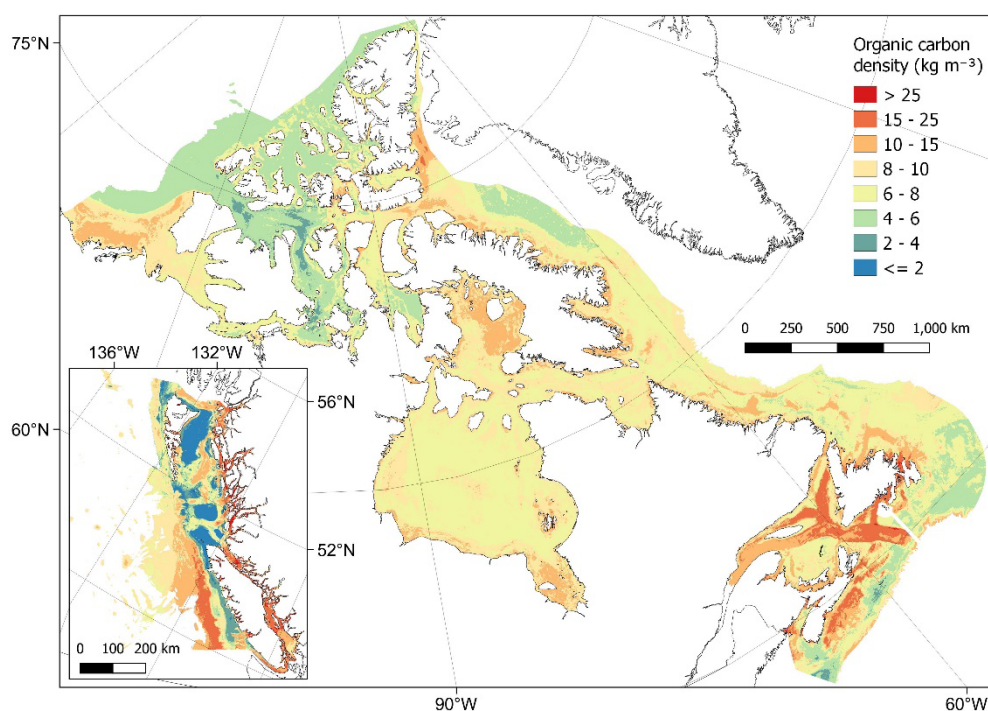
Figure 9. Estimates of sediment dry bulk density (g cm^{-3}) across the Canadian continental margin. The main plot shows the Arctic and Atlantic regions with the Pacific region inset. The estimated bounds of uncertainty around the predicted means are shown in Figure E3. Labels indicating the locations of different areas mentioned within the text are shown in Figure B3. Country outlines from World Bank Official Boundaries, available at https://datacatalog.worldbank.org/search/dataset/0038272.

3.4 Estimated organic carbon density and standing stock

From combining predictions of dry bulk density and organic carbon content, organic carbon density could be estimated across the Canadian continental margin (Fig. 2). Estimated values ranged from 5x10^{-4} to 58.4 kg m^{-3} with a mean of 8.1 ± 2.8 kg m^{-3} (± SD). Spatial patterns in organic carbon density (Fig. 10) were similar to those found for organic carbon content (Fig. 8). Areas with highest carbon density (> 25 kg m^{-3}) were restricted to small areas within nearshore zones, including inlets and fjords of British Columbia (Pacific), as well as enclosed nearshore areas of the Atlantic East Coast (Fig. 10). High carbon densities (> 15 kg m^{-3}) were predicted to occur across wide parts of these areas as well as further offshore in parts of the Laurentian channel and central Scotian Shelf, and at the edge of the continental slope off the West of Vancouver Island (Fig. 10). In the Arctic, areas with relatively high carbon (>10 kg m^{-3}) were predicted across many nearshore areas, as well as across large parts of the Beaufort Shelf, Foxe Basin, James Bay and the Kane Basin (Fig. 10). Cell specific upper and lower uncertainty bounds

701 are shown in Figure E4. On average the upper bounds were 50% higher than the mean prediction,
702 and the lower bounds 37% less than their means.

703 Using a standardised sediment depth of 30 cm, the total standing stock of organic carbon in
704 surficial sediments across the model domain is estimated at 10.9 Gt with uncertainty bounds of
705 7.0 – 16.0 Gt. Between bioregions, total stock was predominantly related to the total areal extent,
706 for example Hudson Bay having the largest carbon stock and largest area (Table 2). The Strait of
707 Georgia and Southern Shelf bioregions of the Pacific had the lowest total standing stocks due
708 their small extent, however per unit area, these regions contained the highest organic carbon
709 stocks, along with the Gulf of St Lawrence.

710

711

712 **Figure 10. Estimates of organic carbon density (kg m⁻³) across the Canadian continental margin.** The main plot
713 shows the Arctic and Atlantic regions with the Pacific region inset. The continuous variable is shown displayed in
714 discrete colour bands to improve visualisation of highly right skewed data. The estimated bounds of uncertainty around
715 the predicted means are shown in Figure E4. Labels indicating the locations of different areas mentioned within the text
716 are shown in Figure B3. Country outlines from World Bank Official Boundaries, available at
717 https://datacatalog.worldbank.org/search/dataset/0038272.

718

Table 2. Summary of estimated mean total organic carbon stocks in surficial seabed sediments of different bioregions across the Canadian continental margin. Organic carbon standing stocks are estimated for the top 30 cm of seabed sediments. For delineation of the different bioregions see Supplement.

Bioregion	Model domain extent (km^2)	OC stock (Gt)	Stock per unit area (kt km^2)
1. Offshore Pacific	53,598	0.15	2.75
2. Northern Shelf BC	96,373	0.23	2.34
3. Southern Shelf BC	28,313	0.10	3.38
4. Strait of Georgia	8,664	0.04	4.94
5. Western Arctic	526,309	1.09	2.06
6. Arctic Basin	250,178	0.42	1.69
7. Arctic Archipelago	243,425	0.47	1.92
8. Eastern Arctic	757,226	1.82	2.40
9. Hudson Bay	1,234,257	3.08	2.49
10. NL Shelves	820,462	2.04	2.49
11. Gulf of St Lawrence	235,541	0.80	3.38
12. Scotian Shelf	234,888	0.65	2.77

Notes: OC = Organic carbon; NL = Newfoundland-Labrador.

3.5 Rock substrate distribution case studies

As the predictive maps produced in this study rely on physical sediment samples alone, they are unlikely to produce valid estimates for areas of bedrock – i.e. estimates of zero sediment carbon density where bedrock is located. On the Scotian shelf (bioregion 12), correcting our predictive maps with a predicted bedrock distribution map (Fig. F1) reduces total organic carbon stock estimates in this region by between 7.7 – 7.8% leading to a value of 0.59 Gt (0.37 – 0.90 Gt). For the Pacific British Columbian marine region (bioregions 1-4), assigning zero values to areas covered by a predicted bedrock distribution map (Fig. F2) would reduce our estimates by 9.1 – 9.7% to a total of 0.46 Gt (0.29 – 0.71 Gt) of organic carbon.

4. Code and data availability

All mapped products as shown in Figures 5, 8, 9 and 10 have been made available as georeferenced TIFF files in the Borealis data repository at https://borealisdata.ca/privateurl.xhtml?token=7bb00f1e-2ce3-400c-955d-e8e0d4fe3080

738 (Epstein et al., 2023). This includes the mean predictions as well as the cell-specific uncertainty
739 bounds as shown in Appendix E. The repository also contains all data collated within the
740 systematic data review of organic carbon content and the georeferenced TIFF files from the rock
741 distribution case studies (Appendix F). Additionally, all the associated code used for data
742 manipulations, model building and predictive mapping can also be found within the above
743 repository.

744

745 **5. Discussion**

746 Using best available data, we have produced the first national assessment of organic carbon in
747 surficial seabed sediments across the Canadian continental margin, estimating the standing stock
748 in the top 30 cm to be 10.9 Gt (7.0 – 16.0 Gt). Although comparisons to previous global studies
749 is challenging due to differences in sediment reference depths, mapping resolutions and total
750 spatial coverage, our estimate falls within a similar range to those previously published (e.g.
751 2.2 Gt in the top 5 cm (Lee et al., 2019) and 48 Gt in the top meter (Atwood et al., 2020) of the
752 Canadian EEZ). In contrast to these global studies, the national approach taken here allows for a
753 more complete data synthesis, a finer spatial resolution, larger spatial coverage of the Canadian
754 continental margin and spatially defined estimates of uncertainty. Similarly to other national and
755 regional mapping studies (Smeaton et al., 2021; Diesing et al., 2017, 2021), areas of high organic
756 carbon stocks were predominantly predicted to occur in coastal fjords, inlets, estuaries, enclosed
757 bays and sheltered basins, as well as in deeper channels and troughs (Fig. 10). To put our
758 estimated organic carbon standing stock into context, 10.9 Gt equates to 52% of the organic
759 carbon estimated to be stored in all Canadian terrestrial plant live biomass and detritus (both
760 above and below ground), and 9.8% of soil organic carbon to 30 cm across Canada (Sothe et
761 al., 2022).

762

763 *Model interpretation and uncertainties*

764 The two key components of the carbon stock estimates in this study are the predictive maps for
765 mud content and organic carbon content, which were estimated to have a map accuracy of 60%
766 and 58% respectively (R^2 0.60 and 0.58). While these values may seem relatively low when
767 compared to some other related studies (Diesing et al., 2017, 2021; Atwood et al., 2020; Mitchell
768 et al., 2019), the use of robust, spatial cross-validation to calculate model evaluation metrics (as

769 we did herein) has been shown to produce significantly more conservative estimates of map

770 accuracy when compared to frequently used random cross-validation approaches (Ludwig et al.,

771 2023; Meyer et al., 2019) such as those used in both the global seabed carbon stock studies

772 discussed above (Atwood et al., 2020; Lee et al., 2019). Within this study, we also calculated cell

773 specific uncertainty bounds. While there are many ways to calculate model uncertainty, therefore

774 making comparisons between studies challenging, the uncertainty in carbon density calculated

775 here (37-50% either side of the mean) is close to those found within similar regional (Diesing et

776 al., 2021; 58%) and global studies (Lee et al., 2019; 49%), both of which predict carbon stocks at

777 significantly coarser resolutions. Our bounds for total standing stock (36% lower and 47% higher

778 than the mean) are also similar to the estimated bounds from the recently published predictive

779 models of Canadian terrestrial vegetation and soil carbon (a 90% confidence interval 48% either

780 side of the mean) (Sothe et al., 2022).

781 Using two case studies from British Columbia and the Scotian Shelf, we estimated that the

782 distribution of rock substrates could reduce our estimates of carbon stock by approximately 7.7 –

783 9.7% (Fig. F1, F2). As much of the Canadian coastline is distant from significant infrastructure,

784 extensive surveys of the seafloor are generally lacking, especially when compared to similar

785 regional carbon mapping studies in northwest Europe (e.g. Smeaton et al., 2021). It is therefore

786 unclear how representative these case studies are of the entire Canadian EEZ. Improved data on

787 the presence of bedrock across lesser studied regions of the Canadian Arctic, Hudson Bay, Gulf

788 of St Lawrence, Newfoundland and Labrador may allow for the production of a predictive map of

789 bedrock across the Canadian EEZ which would significantly improve the carbon estimates and

790 spatial predictive maps produced in this study.

791 Areas of uncertainty which could not be fully quantified include the accuracy and precision of

792 response data and predictor layers. The response data drive the model construction, and

793 therefore sampling, processing, or recording errors can propagate into predictions. This is

794 particularly relevant given the large temporal extent of response data which was required to gain

795 sufficient coverage for this work (1959-2019). This large duration may also add additional variation

796 from temporal differences between data, for example from differing anthropogenic drivers on

797 carbon storage and/or accumulation (Keil, 2017); however, similar temporal extents have been

798 used in related studies (Atwood et al., 2020; Lee et al., 2019; Seiter et al., 2004) and 72% of the

799 organic carbon data within this study were sampled after 1980 and 55% after 2000. Within the

800 response data, assumptions and/or predictions were also required regarding the distribution of

801 mud and carbon across sediment depths. While standardising for this factor is clearly necessary,

802 especially when using a wide variety of legacy data, it does add additional uncertainty which would
803 not be present if widescale standardised sampling methods were employed. The results from this
804 study do however highlight, that within the top 30 cm of sediment, the spatial location of the
805 sample is a far stronger driver of organic carbon content than the sediment sampling depth (Table
806 C1).

807 Most of the predictor variables used in this study are also themselves modelled products, which
808 contain their own inherent uncertainties and/or interpolations which cannot be fully quantified
809 here. Additionally, many predictors are constructed at spatial resolutions significantly coarser than
810 that used for modelling and prediction in this study. This meant that predictor data had to be
811 interpolated, with significant inherent assumptions regarding variation and distribution of the data.
812 Although best available data were used in this study, if predictor variables were available at higher
813 native resolutions, less assumptions would be necessary and significant differences may be found
814 in predictions, as well as their uncertainty and variability. Many of the predictor variables also
815 have temporal components, and while the climatological mean of a 12 – 14 year timespan used
816 in this study is expected to produce variables representative for the study region, they do not
817 completely align with the temporal extent of the response data which could add further prediction
818 uncertainty. Finally, due to data availability, the uncertainty bounds around our mean estimates
819 of dry bulk density and organic carbon density were approximated from the constructed 95% CIs
820 of mud content and %OC from the random forest models. While these provide an appropriate
821 measure of uncertainty in our estimates in the context of this study, if large empirical datasets
822 became available for dry bulk and organic carbon density, it would be preferable to construct
823 predictive models, mean estimates and uncertainty bounds for these response variables directly.

824

825 *Future directions and applications*

826 Improvements could be made in future iterations of these sediment carbon maps when additional
827 response data become available. The size of the organic carbon content dataset was relatively
828 small (2,518 point-samples) given the size of the model domain, so new data could greatly
829 improve accuracy and reduce uncertainty in predictions. Additionally, wide-spread empirical data
830 on sediment dry bulk density would reduce the assumptions needed in using approximate
831 conversions from mud content, while a large geographically dispersed empirical dataset on
832 seabed sediment organic carbon density (i.e. where OC content and dry bulk density is measured
833 directly in each physical sample) would reduce assumptions even further, with the potential to

834   construct a single predictive map for this response alone (Diesing et al., 2021). There are also
835   improvements to be made with the development of higher resolution or more accurate predictor
836   layers. This would be particularly relevant for those variables with coarse resolutions and those
837   which were seen to have highest importance within our models or from related seabed sediment
838   mapping studies (e.g. Gregr et al., 2021; Diesing et al., 2017, 2021; Mitchell et al., 2019) – i.e.
839   wave velocities, suspended particulate matter, exposure, current velocities and oxygen
840   concentrations. Further validation and refinements could also be supported by numerical
841   biogeochemical modelling products where the organic carbon densities are mathematically
842   estimated based on oceanographic, climatological and benthic conditions, including the potential
843   to incorporate predictions under different future climate scenarios (Ani and Robson, 2021).

844   The organic carbon predictive mapping product generated here could have many future
845   applications. Regionalisation and prioritisation processes could identify key areas of carbon
846   storage for further research and possible protections (Epstein and Roberts, 2022, 2023; Diesing
847   et al., 2021). There is also potential to combine these mapped products with spatial data on
848   human activities occurring on the seafloor to consider potential management implications, such
849   as controlling the levels of impactful industries (e.g. mobile bottom fishing, mineral extraction,
850   energy generation) in areas with high organic carbon (Clare et al., 2023; Epstein and Roberts,
851   2022). The mud content predictive maps may also have applications for marine planning more
852   widely, being a strong driver of the biological habitat type and sensitivity. Overall, these data have
853   wide-scale relevance across marine ecology, geology and environmental management
854   disciplines, however, the use of these products should always consider the discussed
855   uncertainties and quantified uncertainty bounds of predictions. As with all large-scale mapping
856   exercises, continued empirical data collection is needed for improved accuracy of mapping
857   seabed carbon stocks across Canada.

6. Appendices

Appendix A. Supplementary methods

A1. Analysis software

Analyses were primarily undertaken in R 4.2.2 (R Core Team, 2022) and Rstudio 2022.12.0.353 (Posit Team, 2022), with some additional data manipulation and spatial plotting in QGIS (QGIS.org, 2021) and Python (Van Rossum and Drake, 2009). Within R, raster data were handled using the *terra* package (Hijmans, 2022), spatial vector data using the *sf* package (Pebesma, 2018), netCDF data with the *stars* (Pebesma, 2022) and *tidync* (Sumner, 2022) packages, data-frames with the *dplyr* package (Wickham et al., 2019), and vector data with base R (R Core Team, 2022). Random forest modelling was primarily dependent on the *ranger* package (Wright and Ziegler, 2017), however models were constructed and tuned using the *tidymodels* package (Kuhn and Wickham, 2020), with cross-validation and predictor variable selection using the *CAST* (Meyer et al., 2023) and *caret* (Kuhn, 2022) packages. Plotting utilised the above packages as well as *ggplot2* (Wickham et al., 2019) and *patchwork* (Pedersen, 2022) while parallel processing used the *doParallel* package (Microsoft Corporation and Weston, 2022).

A2. Bathymetry layer construction

To define the maximum potential spatial coverage of this study, best available bathymetric datasets were combined across the Canadian Exclusive Economic Zone (EEZ) (Table 1). Firstly, three Digital Elevation Model (DEM) raster layers covering different extents of the Canadian EEZ were each filtered to contain only those elevations of less than or equal to 0 m. Where necessary, data were then aggregated (averaged) or disaggregated (split) to a resolution of approximately 200 m, and all layers were projected onto a unified 200 m x 200 m equal area grid (co-ordinate reference system (CRS) EPSG:3573 – WGS 84 – North Pole Lambert Azimuthal Equal Area Canada). Reprojection was necessary as all three DEMs were in different co-ordinate systems, including some already being projected. The 200 m resolution was chosen as it is the median native resolution of the three DEMs, while also being considered towards the upper limit of what may be computationally possible within the scope of this study. After reprojection, the three layers were overlain, with the region-specific data given priority over global data where present. Finally, the seaward boundaries were delineated by the outer extent of the Canadian EEZ (Flanders Marine Institute, 2019).

889

A3. Details of ocean circulation models

ANHA12 is a regional configuration of the NEMO ocean and sea-ice model (Madec et al., 1998) created at the University of Alberta, covering the Arctic and northern Hemisphere Atlantic at 5 day temporal resolution, a curvilinear 1/12th degree horizontal resolution ranging from 1.93 km in the Arctic to 9.3 km at the equator, and 50 vertical levels (Hu et al., 2019). The British Columbia continental margin (BCCM) circulation model created by Fisheries and Oceans Canada (DFO) covers the entire Canadian Pacific coast and extends approximately 400 km offshore. It has a uniform horizontal resolution of 3 km, 42 vertical levels and a 3 day temporal resolution (Peña et al., 2019; Masson and Fine, 2012). As the BCCM model has higher uncertainty in nearshore and enclosed environments due to its relatively coarse resolution, data were also extracted for the enclosed Salish Sea from the Salish Sea Cast ERDDAP data server. Similarly to the ANHA12 model, the Salish Sea Cast is a configuration of the NEMO circulation model developed by a consortium of Canadian Universities and government agencies and extends from Juan de Fuca Strait to Puget Sound to Johnstone Strait at 500 m horizontal resolution, 40 vertical layers and hourly temporal resolution (Soontiens and Allen, 2017; Soontiens et al., 2016). For further details on all these models, see relevant cited references. It should be noted that many of these ocean circulation models contain high uncertainty in nearshore areas. However, they are expected to be greatly improved when compared to global circulation model products (Peña et al., 2019; Hu et al., 2019; Soontiens and Allen, 2017) which are frequently used in this sort of predictive mapping work (e.g. Atwood et al., 2020; Lee et al., 2019; Assis et al., 2018).

910

A4. Sediment grain size data collation and processing details

Sediment composition point data were extracted from two sources. Firstly, all data were exported from the NRCan Expedition Database on 11th November 2022. This data repository contains information related to marine and coastal field surveys conducted by or on behalf of the Geological Survey of Canada from the 1950s to present, which deployed sampling methods including piston cores and grab samples. Data were also extracted from a recent synthesis of grain size distribution measurements from the Canadian Pacific seafloor (1951-2017), compiled by Geological Survey Of Canada and NRCan (Enkin, 2023). Although there are some duplications between these two datasets, these are accounted for in proceeding pre-processing steps. In both sources, grain size data is reported as the percentage content of mud (sometimes separated into

921 silt and clay), sand and gravel within each sample. Due to modern developments in grain size
922 analyses (e.g. laser diffraction) older samples may have lower measurement accuracy; however,
923 due to the relatively coarse metric being used in this study (%mud/sand/gravel) and the
924 occurrence of a number of largescale geological surveys occurring during the 1960s, we chose
925 to retain data from 1960 onwards. Where sampling year was not recorded within the database,
926 the date was inferred from the expedition code or from expedition metadata. The sampling method
927 and depth of the sediment from which the sample/sub-sample originates are also predominantly
928 recorded within the database. Where sediment depth was absent, but the sampling method was
929 noted as "grab" or "other", the penetration depth was assumed to be 10 cm (a commonly assumed
930 penetration of standard sediment sampling devices such as Van Veen Grabs and Day Grabs).

931

932 A5. Details on construction and implementation of spatial cross validation and feature selection

933 For each response variable modelled in this study (mud content and organic carbon content
934 (%OC)), the *spatialsample* package (Silge and Mahoney, 2023) was used to construct a variety
935 of spatial CV data-fold structures and the validity of each structure was visually assessed using
936 the *CAST.plot_geodist* function (Meyer et al., 2023). This function creates density plots of nearest
937 neighbour distances (Euclidean) in multivariate predictor space (using normalized variables)
938 between response data locations and a random sample of prediction locations, and between data
939 inside and outside each CV fold (Ludwig et al., 2023; Meyer et al., 2023; Meyer and Pebesma,
940 2022). The suitability of a given CV structure to be representative of estimating map accuracy can
941 be determined by visually assessing the density plots and finding the CV-distance curve being
942 closely aligned to the density curve of response data to prediction distances (see Appendix D;
943 Ludwig et al., 2023; Meyer and Pebesma, 2022). To approximate response-to-prediction
944 distances, the sample size number within *plot_geodist* was set to select 5,000 random samples
945 across the model domain. Further, as the spatial distribution of data is a key consideration to
946 ensure robust cross-validation (Ludwig et al., 2023; Meyer and Pebesma, 2022), for the
947 *plot_geodist* calculations alone, the x- and y-coordinates of each data point were included in
948 addition to those predictor variables listed in Table 1 and described in Section 2.5.

949 For the mud content data, a spatial kmeans clustering CV structure was chosen as the response
950 data had good coverage of the model domain, contained a large number of data points, and
951 showed relatively strong spatial clustering (Fig. B1). A range of options in the number of kmeans
952 clusters were tested, with 35 being determined as the optimal number and each cluster being

953   assigned to its own CV fold (Fig. D1). Through visual assessment of the density plots, it was
954   identified that the kmeans CV structure was somewhat mis-aligned from response-to-prediction
955   distances, with the CV distances being overly conservative at including near-distance
956   comparisons (Fig. D1). We therefore used a partially repeated CV strategy, with a small number
957   of randomly selected data-points added to the assessment set in each kmeans spatial-CV fold
958   (1% of mud content data randomly sampled at each fold without replacement) (Fig. D2). As the
959   %OC response dataset was relatively small and spatially dispersed (Fig. B2), we used a spatial
960   block CV strategy in place of the kmeans clustering to avoid clusters containing highly spatially
961   dispersed data. We chose to use hexagonal shaped blocks, random assignment of blocks to folds,
962   and the same number of CV folds as for the mud content data (v = 35) – both to maintain
963   uniformity and because varying the fold-number did not significantly influence the density plots.
964   Instead, the diameter of the spatial blocks was altered, and an optimal block size of 100 km
965   identified using the *plot_geodist* function (Fig. D3).

966   The *CAST.ffs* function (Meyer et al., 2023) was used to run a forward predictor variable selection
967   process with appropriate spatial considerations. The function fits a model with all combinations of
968   two-way predictors, selects the best model based on a given metric, and then increases the
969   number of predictors by one, testing all remaining variables. This iteratively continues with the
970   process stopping if none of the tested variables increases the performance when compared to
971   the best previous model with "n-1" predictors. The function allows models to be fit separately on
972   each spatial CV fold (as defined above),, with the overall performance of each iteration based on
973   model accuracy across all CV folds. This therefore incorporates appropriate spatial considerations
974   into the feature selection process. Due to the large number of variables within this study, and the
975   relatively large datasets, this process was very computationally expensive. We therefore chose
976   to adapt the function to initiate forward variable selection after initial identification of the first two
977   predictor variables. These variables were identified by constructing a single random forest model
978   with all training data and predictor variables, and the hyperparameters *mtry* (the number of
979   variables to randomly sample as candidates at each split), *min_n* (the number of observations
980   needed to keep splitting nodes) and *trees* (the number of random forest trees to construct and
981   take mean predictions across) set to 2, 5 and 1,000 respectively. Variable importance was
982   estimated on out-of-bag samples through permutation of predictor variable values (Wright et al.,
983   2016), and the two predictor variables with highest importance selected. The *ffs* function was then
984   run starting with the two pre-selected variables (see Fig. 3 & 6) and performance of each iteration
985   assessed on the root mean squared error (RMSE) of predictions across all CV folds.

986

Appendix B. Distribution of response data

988

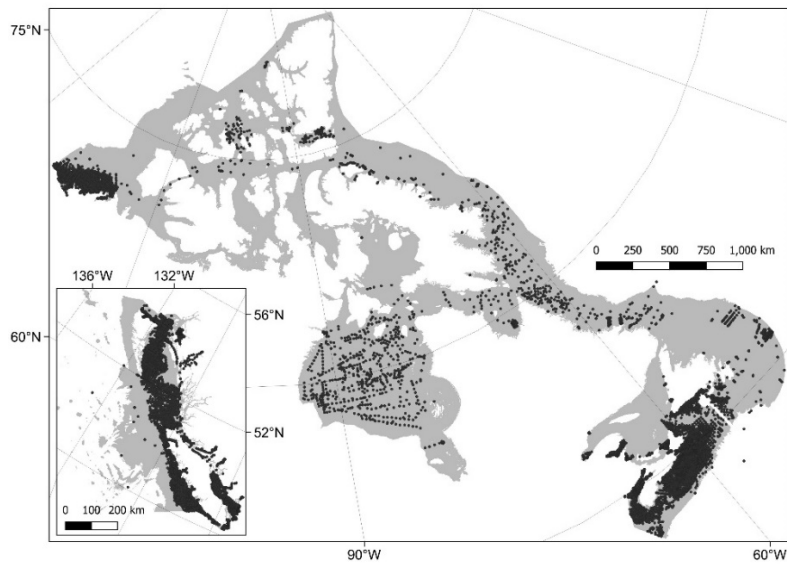



Figure B1. Map showing the distribution of mud content samples across the model domain.

990

991

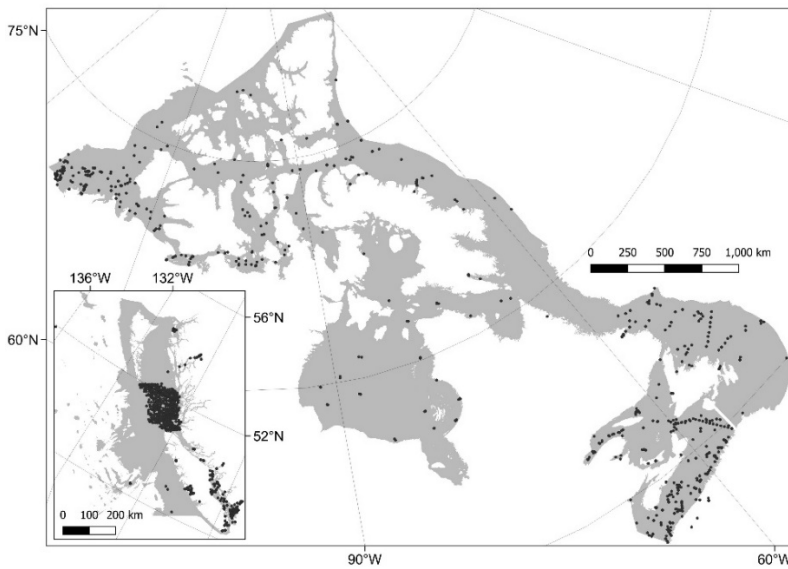



Figure B2. Map showing the distribution of carbon content samples across the model domain.

993

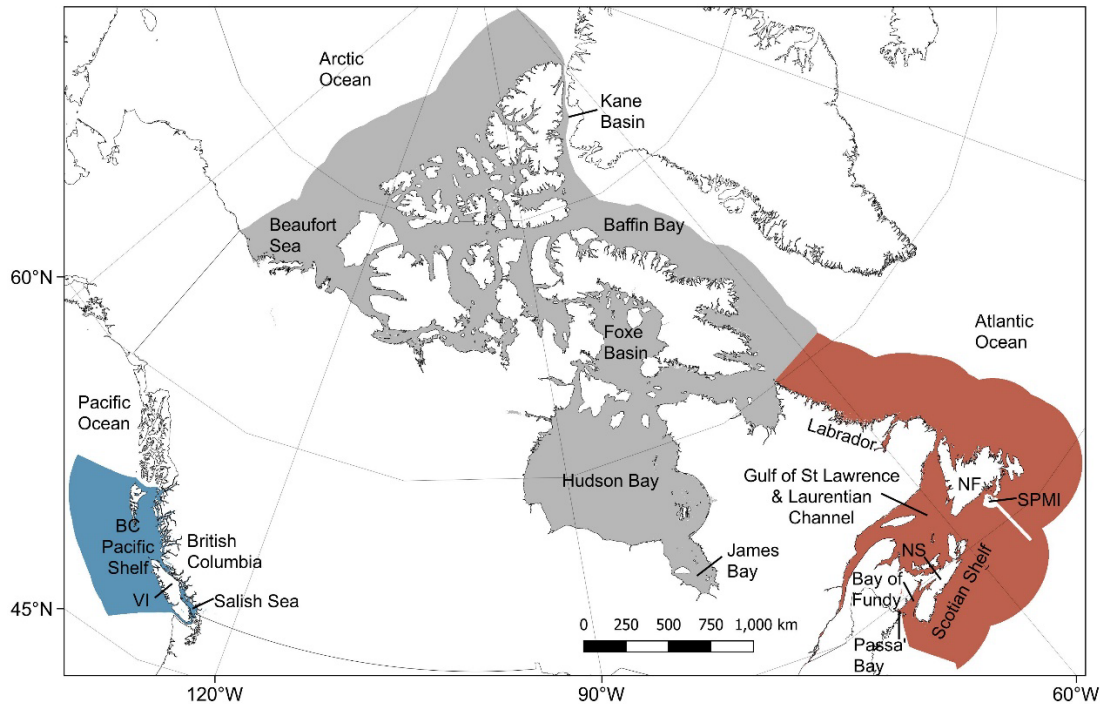



Figure B3. Map indicating the locations of different areas which are mentioned within the text. The Canadian Pacific (blue), Arctic (grey) and Atlantic (red) regions are shown with labelled locations overlayed. BC = British Columbia; Passa' Bay = Passamaquoddy Bay; NS = Nova Scotia; NF = Newfoundland; SPMI = St Pierre and Miquelon. The locations are for guidance only and do not represent the entire extent or exact location of a given area. Country outlines are derived from World Bank Official Boundaries, available at https://datacatalog.worldbank.org/search/dataset/0038272.

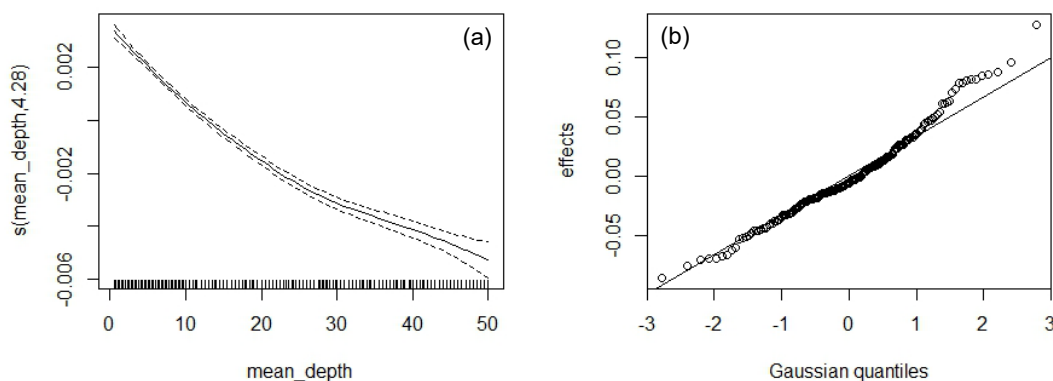
Appendix C. Organic carbon sediment depth modelling results

There was a significant effect of sampling depth on the organic carbon content in seabed sediments ($\chi^2 = $ 1400.9, $p < 0.001$). While sample ID explained most of the variation between sub-sample carbon contents, the sampling depth was also a significant factor (Table C1). Carbon content decreased with increasing sampling depth (Fig. C1). The rate of carbon content decline generally decreased with increasing depth into the sediment, however uncertainty in this trend increased within deeper sediment layers (Fig. C1).

1012 **Table C1. Results from the generalised additive mixed model between the carbon content of marine sediments**
1013 **and sampling depth**. A basic generalised additive mixed model with a scaled-t distribution was constructed for carbon
1014 content in sediment sub-samples with sample ID as the random factor and sampling depth as the fixed factor.

Spline	Type	edf	Res. df	χ^2	Deviance explained	p
Sampling depth (cm)	Cubic	4.28	5.36	2299	1.1%	< 0.001
ID	Random	181.94	182.00	715046	86.9%	< 0.001

1015 Notes: edf = Effective degrees of freedom. Res. df = Residual degrees of freedom

1016

1017

1018 **Figure C1. Regression splines indicating the effect of sediment sampling depth (a) and sample ID (b) on the**
1019 **organic carbon content in seabed sediment sub-samples.**

1020

1021 The predicted mean effect of sediment depth on carbon content was extracted from the model
1022 and converted into a transfer function which states the expected ratio between the carbon content
1023 across 30 cm compared to the cumulative mean at any given sampling depth (Figure C2). The
1024 ratio ranged from 89.3% when only measuring the sediment surface, to 93.7% if measuring the
1025 carbon content across the top 10 cm, and by 25 cm was approaching equilibrium at 98.8%.

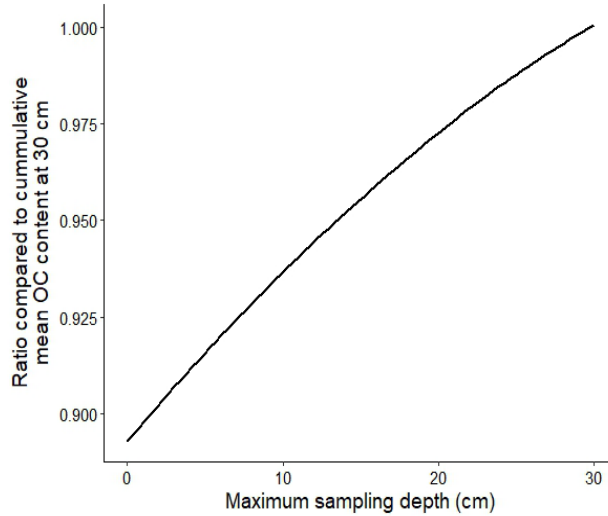
1026

1027

Figure C2. Transfer function for cumulative mean organic carbon (OC) content at 30 cm sediment depth. Using a generalised additive mixed model an estimated transfer function was constructed to standardise the cumulative mean carbon content at any given depth to an expected value at 30 cm.

1028
1029
1030

1031

Appendix D. Results from random forest cross-validation structure selection

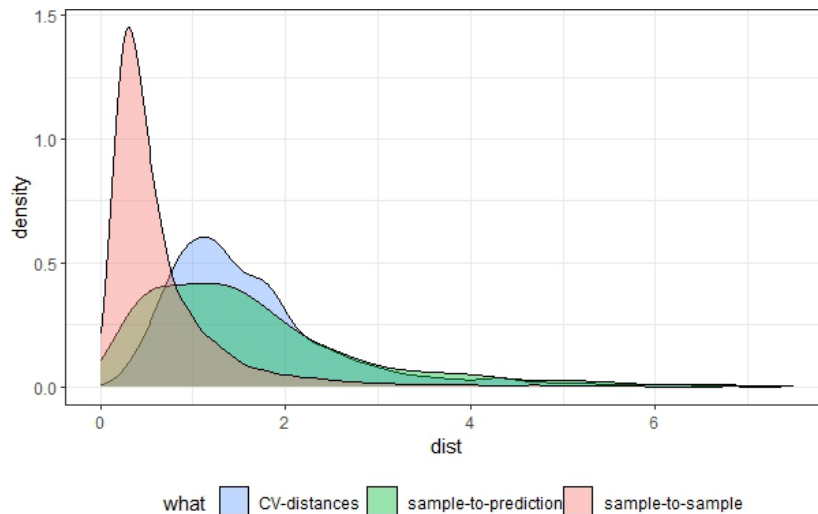
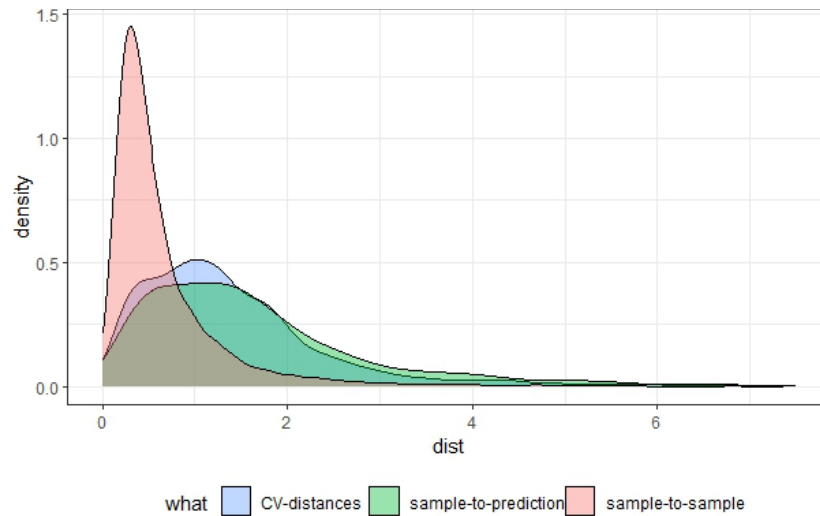
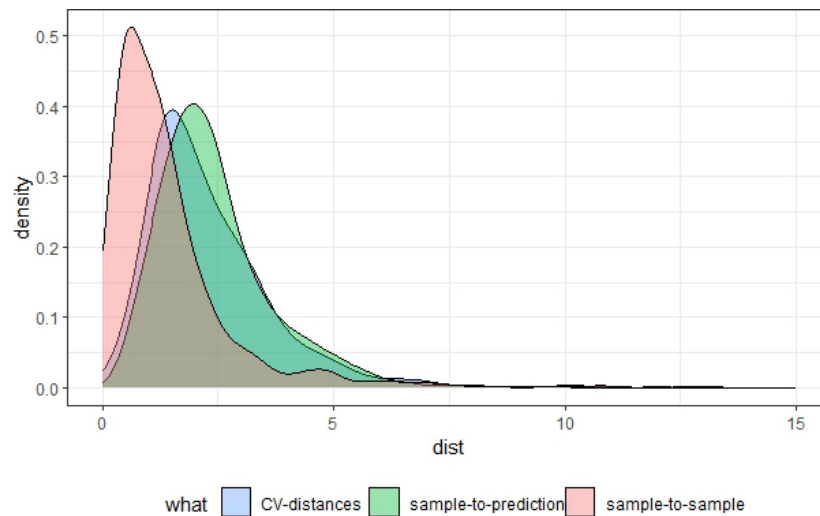
1032

1033

Figure D1. Multivariate nearest-neighbour distance density plot for mud content data with the optimal number of spatial k-means clusters across cross validation (CV) folds. Frequency of nearest neighbour distances (x-axis) is shown for sample-to-sample distance (red), sample-to-prediction distance (green) and CV fold analysis-to-assessment distance (blue). dist = Multivariate Euclidean distance in predictor space after normalization of predictors. An optimal number of 35 clusters was selected to due close overlap between the CV-distance and sample-to-prediction curve.

1034
1035
1036
1037
1038
1039

1040

1041 **Figure D2. Multivariate nearest-neighbour distance density plot for mud content data with a partially repeated**
1042 **spatial-random mixture method for cross validation (CV) folds.** Frequency of nearest neighbour distances (x-axis)
1043 is shown for sample-to-sample distance (red), sample-to-prediction distance (green) and CV fold analysis-to-
1044 assessment distance (blue). dist = Multivariate Euclidean distance in predictor space after normalization of predictors.
1045 Due to the optimal spatial k-means clustering showing poor overlap at lower multivariate distances (Fig. D1), a 1%
1046 random sample without replacement was added to each fold.

1047

1048

1049 **Figure D3. Multivariate nearest neighbour distance density plot for organic carbon content data with the**
1050 **optimal block size across cross validation (CV) folds.** Frequency of nearest neighbour distances (x-axis) is shown
1051 for sample-to-sample distance (red), sample-to-prediction distance (green) and CV fold analysis-to-assessment
1052 distance (blue). dist = Multivariate Euclidean distance in predictor space after normalization of predictors. An optimal
1053 block size of 100 km was selected to due close overlap between the CV-distance and sample-to-prediction curve.

1054

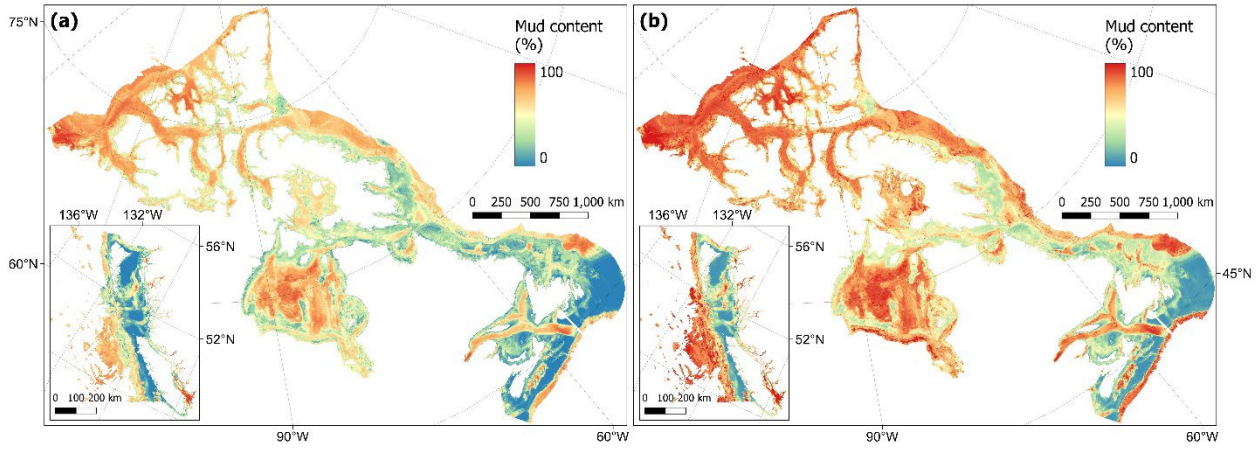
 Appendix E. Cell-specific uncertainty bounds for predictive sediment maps

1056

Figure E1. Estimated lower (a) and upper (b) bounds of the 95% confidence interval for predictions of mud content (%) in subtidal marine sediments across the Canadian continental margin. Within each panel the main plot shows the Arctic and Atlantic regions with the Pacific region inset.

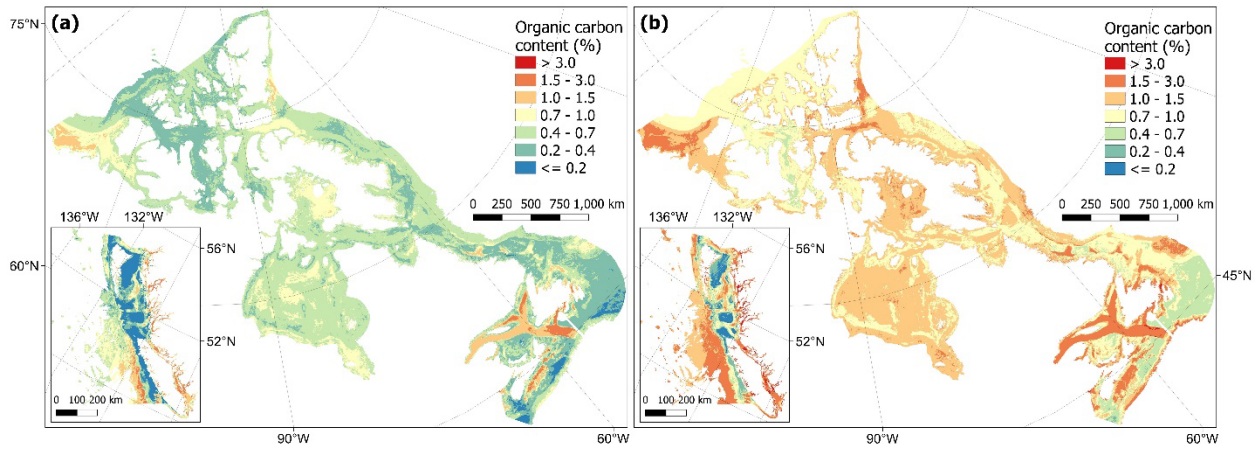
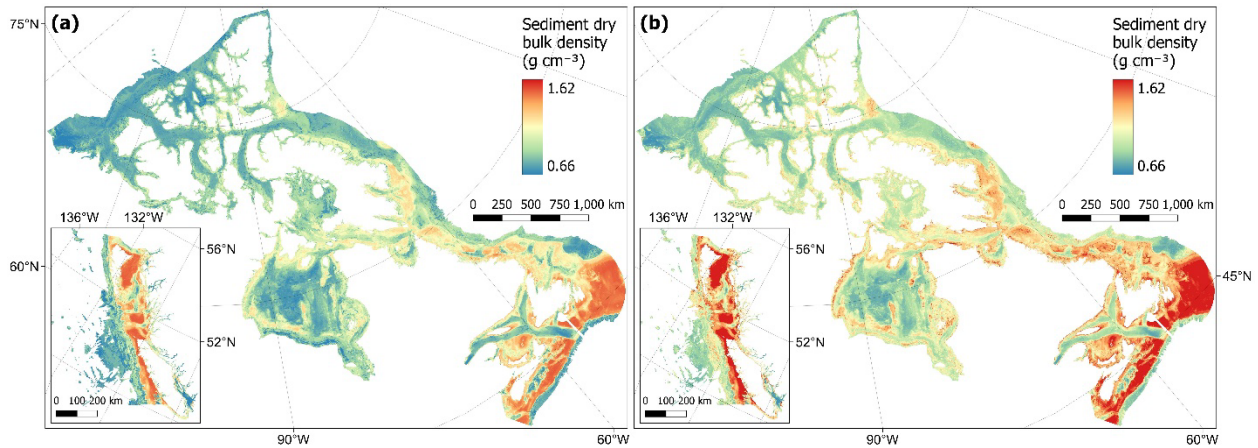
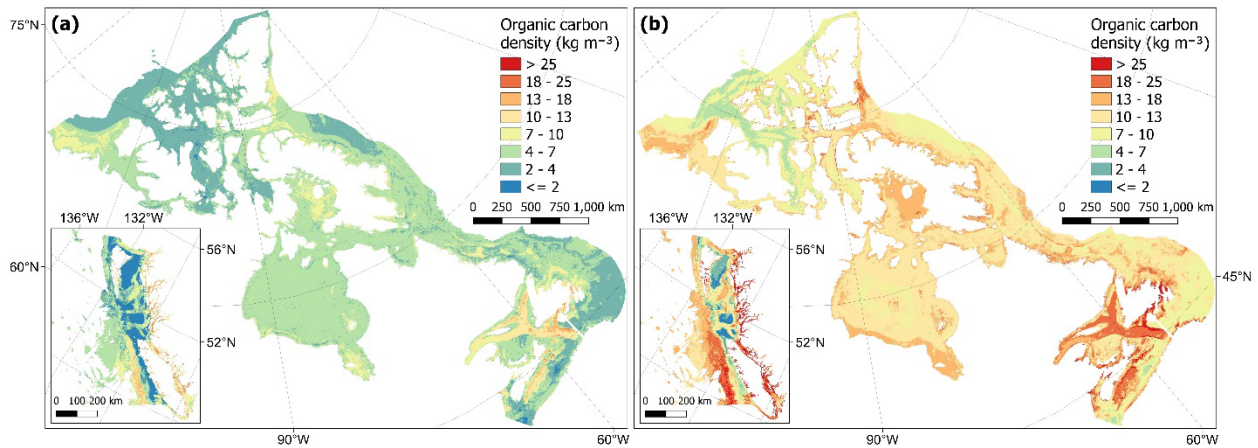
1060

1061

Figure E2. Estimated lower (a) and upper (b) bounds of the 95% confidence interval for predictions of carbon content (%) in subtidal marine sediments across the Canadian continental margin. The continuous variable is shown in discrete colour bands to improve visualisation of highly right skewed data. Within each panel the main plot shows the Arctic and Atlantic regions with the Pacific region inset.

1066

1067

1068

1069 **Figure E3. Estimated lower (a) and upper (b) uncertainty bounds around the mean predictions of dry bulk**
1070 **density (g cm⁻³) of subtidal marine sediments across the Canadian continental margin.** Within each panel the
1071 main plot shows the Arctic and Atlantic regions with the Pacific region inset.

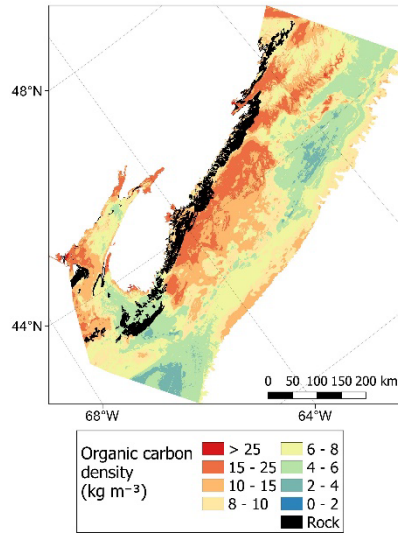
1072

1073

1074 **Figure E4. Estimated lower (a) and upper (b) uncertainty bounds around the mean predictions of organic**
1075 **carbon density (kg m⁻³) in subtidal marine sediments across the Canadian continental margin.** The continuous
1076 variable is shown in discrete colour bands to improve visualisation of highly right skewed data. Within each panel the
1077 main plot shows the Arctic and Atlantic regions with the Pacific region inset.

1078

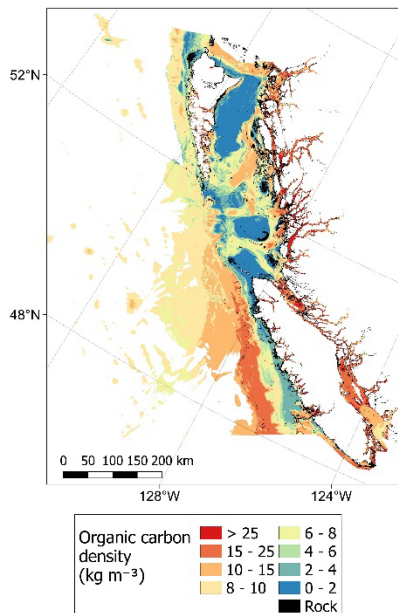
1079

1080

1081   **Appendix F. Bedrock distribution case studies**

1082

1083   **Figure F1. Predicted mean values of organic carbon density within the Scotian Shelf overlayed by the estimated**
1084   **distribution of rock substrates.** Data on the estimated distribution of rock on the seafloor across the Scotian Shelf
1085   Bioregion is taken from Philibert et al. (2022).

1086

1087

1088

1089   **Figure F2. Predicted mean values of organic carbon density within the British Columbia EEZ overlayed by the**
1090   **estimated distribution of rock substrates.** Data on the estimated distribution of rock on the seafloor across the British
1091   Columbian continental margin is taken from Gregr et al. (2021).

7. Author contributions

JKB and SDF secured funding and led the management of this project. GE, SDF and JKB conceptualised this study. GE, DH, AP, CP & PGM collated the data. GE developed the model code and performed the investigations with input from SDF and JKB throughout. GE prepared the manuscript with contributions from all co-authors.

8. Competing interests

The authors declare that they have no conflict of interest.

9. Acknowledgements

We greatly appreciate advice given across various stages of the processes from Cooper Stacey, Markus Diesing, Sophia Johannessen, Ashley Park, Nadja Steiner, Diane Lavoie, Amber Holdsworth, Michael Li, Kate Jarret, Javier Murillo-Perez, Ellen Kenchington, Emily Rubidge and others within the Department for Fisheries and Oceans, and Natural Resources Canada. We would also like to thank Randy Enkin, Sarah Paradis and Genevieve Philibert for sharing data towards this work, as well as the constructive and insightful comments from the two referees, and Jennifer McHenry, Matt Csordas and Brian Timmer for their ideas towards troubleshooting during the project. This research was enabled in part by support provided by BC & Prairies Digital Research Infrastructure and the Digital Research Alliance of Canada (alliancecan.ca).

10. Financial support

This work was funded by an Natural Sciences and Engineering Research Council (NSERC) Alliance partnership grant #ALLRP571068 – 21 to JKB, and is publication #001 of Blue Carbon Canada. GE is also supported by a Mitacs-Accelerate Fellowship, jointly funded by Oceans North. PGM also gratefully acknowledge the financial and logistic support of grants from NSERC including a Discovery Grant (rgpin 227438-09) and Climate Change and Atmospheric Research Grants (VITALS - RGPCC433898 and the Canadian Arctic Geotraces program - RGPCC 433848), as well as support from the Marine Environmental Observation, Prediction and Response Network (MEOPAR) Prediction Core.

11. References

Amoroso, R. O., Pitcher, C. R., Rijnsdorp, A. D., McConnaughey, R. A., Parma, A. M., Suuronen, P., Eigaard, O. R., Bastardie, F., Hintzen, N. T., Althaus, F., Baird, S. J., Black, J., Buhl-Mortensen, L., Campbell, A. B., Catarino, R., Collie, J., Cowan, J. H., Jr., Durholtz, D., Engstrom, N., Fairweather, T. P., Fock, H. O., Ford, R., Galvez, P. A., Gerritsen, H., Gongora, M. E., Gonzalez, J. A., Hiddink, J. G., Hughes, K. M., Intelmann, S. S., Jenkins, C., Jonsson, P., Kainge, P., Kangas, M., Kathena, J. N., Kavadas, S., Leslie, R. W., Lewis, S. G., Lundy, M., Makin, D., Martin, J., Mazor, T., Gonzalez-Mirelis, G., Newman, S. J., Papadopoulou, N., Posen, P. E., Rochester, W., Russo, T., Sala, A., Semmens, J. M., Silva, C., Tsolos, A., Vanelslander, B., Wakefield, C. B., Wood, B. A., Hilborn, R., Kaiser, M. J., and Jennings, S.: Bottom trawl fishing footprints on the world's continental shelves, Proceedings of the National Academy of Sciences of the USA, 115, E10275–E10282, https://doi.org/10.1073/pnas.1802379115, 2018.

Ani, C. J. and Robson, B.: Responses of marine ecosystems to climate change impacts and their treatment in biogeochemical ecosystem models, Marine Pollution Bulletin, 166, 112223, https://doi.org/10.1016/j.marpolbul.2021.112223, 2021.

Arndt, S., Jørgensen, B. B., LaRowe, D. E., Middelburg, J. J., Pancost, R. D., and Regnier, P.: Quantifying the degradation of organic matter in marine sediments: A review and synthesis, Earth-Science Reviews, 123, 53–86, https://doi.org/10.1016/j.earscirev.2013.02.008, 2013.

Assis, J., Tyberghein, L., Bosch, S., Verbruggen, H., Serrão, E. A., and De Clerck, O.: Bio-ORACLE v2.0: Extending marine data layers for bioclimatic modelling, Global Ecology and Biogeography, 27, 277–284, https://doi.org/10.1111/geb.12693, 2018.

Atwood, T. B., Witt, A., Mayorga, J., Hammill, E., and Sala, E.: Global Patterns in Marine Sediment Carbon Stocks, Frontiers in Marine Science, 7, 165, https://doi.org/10.3389/fmars.2020.00165, 2020.

Avelar, S., van der Voort, T. S., and Eglinton, T. I.: Relevance of carbon stocks of marine sediments for national greenhouse gas inventories of maritime nations, Carbon Balance and Management, 12, 10, https://doi.org/10.1186/s13021-017-0077-x, 2017.

Bauer, J. E., Cai, W.-J., Raymond, P. A., Bianchi, T. S., Hopkinson, C. S., and Regnier, P. A. G.: The changing carbon cycle of the coastal ocean, Nature, 504, 61–70, https://doi.org/10.1038/nature12857, 2013.

Berner, R. A.: Burial of organic carbon and pyrite sulfur in the modern ocean; its geochemical and environmental significance, American Journal of Science, 282, 451–473, https://doi.org/10.2475/ajs.282.4.451, 1982.

Burdige, D. J.: Preservation of Organic Matter in Marine Sediments: Controls, Mechanisms, and an Imbalance in Sediment Organic Carbon Budgets?, Chemical Reviews, 107, 467–485, https://doi.org/10.1021/cr050347q, 2007.

Cavan, E. L. and Hill, S. L.: Commercial fishery disturbance of the global ocean biological carbon sink, Global Change Biology, 28, 1212–1221, https://doi.org/10.1111/gcb.16019, 2022.

Clare, M. A., Lichtschlag, A., Paradis, S., and Barlow, N. L. M.: Assessing the impact of the global subsea telecommunications network on sedimentary organic carbon stocks, Nature Communications, 14, 2080, https://doi.org/10.1038/s41467-023-37854-6, 2023.

1159   Copernicus: Arctic Ocean Wave Hindcast - ARCTIC_MULTIYEAR_WAV_002_013 - Norwegian
1160   Meteorological Institute, Copernicus Marine Data Store, https://doi.org/10.48670/moi-00008, 2022a.

1161   Copernicus: Global Ocean Colour (GlobColour) - ACRI - OCEANCOLOUR_GLO_BGC_L3_MY_009_103 -
1162   Bio-Geo-Chemical, L3 (daily) from Satellite Observations (1997-ongoing), Copernicus Marine Data Store,
1163   https://doi.org/10.48670/moi-00280, 2022b.

1164   Copernicus: Global Ocean Waves Reanalysis - WAVERYS - GLOBAL_MULTIYEAR_WAV_001_032 -
1165   Mercator Océan International, Copernicus Marine Data Store, https://doi.org/10.48670/moi-00022,
1166   2022c.

1167   DFO: Federal Marine Bioregions, Fisheries and Oceans Canada, 2022.

1168   Diesing, M., Kroger, S., Parker, R., Jenkins, C., Mason, C., and Weston, K.: Predicting the standing stock of
1169   organic carbon in surface sediments of the North-West European continental shelf, Biogeochemistry,
1170   135, 183–200, https://doi.org/10.1007/s10533-017-0310-4, 2017.

1171   Diesing, M., Thorsnes, T., and Bjarnadóttir, L. R.: Organic carbon densities and accumulation rates in
1172   surface sediments of the North Sea and Skagerrak, Biogeosciences, 18, 2139–2160,
1173   https://doi.org/10.5194/bg-18-2139-2021, 2021.

1174   Diesing, M., Paradis, S., Jensen, H., Thorsnes, T., Bjarnadóttir, L. R., and Knies, J.: Organic Carbon Stocks
1175   and Accumulation Rates in Surface Sediments of the Norwegian Continental Margin, ESS Open Archive,
1176   pre-print, 2023.

1177   Duarte, C. M., Middelburg, J. J., and Caraco, N.: Major role of marine vegetation on the oceanic carbon
1178   cycle, Biogeosciences, 2, 1–8, 2005.

1179   Enkin, J. R.: Sediment Grain Size Distribution Measurements, from Canadian Pacific Seafloor Samples,
1180   Collected from 1951 to 2017, Geological survey of Canada Open File, in press, 2023.

1181   Epstein, G. and Roberts, C. M.: Identifying priority areas to manage mobile bottom fishing on seabed
1182   carbon in the UK, PLOS Climate, 1, e0000059, https://doi.org/10.1371/journal.pclm.0000059, 2022.

1183   Epstein, G. and Roberts, C. M.: Does biodiversity-focused protection of the seabed deliver carbon
1184   benefits? A U.K. case study, Conservation Letters, 16, e12929, https://doi.org/10.1111/conl.12929,
1185   2023.

1186   Epstein, G., Middelburg, J. J., Hawkins, J. P., Norris, C. R., and Roberts, C. M.: The impact of mobile
1187   demersal fishing on carbon storage in seabed sediments, Global Change Biology, 28, 2875–2894,
1188   https://doi.org/10.1111/gcb.16105, 2022.

1189   Epstein, G., Fuller, S. D., Hingmire, D., Myers, P., Peña, A., Pennelly, C., and Baum, J. K.: Predictive maps
1190   and related data on organic carbon stocks and accumulation rates in surficial sediments of the Canadian
1191   continental margin. https://borealisdata.ca/privateurl.xhtml?token=7bb00f1e-2ce3-400c-955d-
1192   e8e0d4fe3080, Borealis, DRAFT VERSION, UNF:6:vYJjKEPuOf4yOyxq1b4dFg== [fileUNF], 2023.

1193   Evans, J. S. and Murphy, M. A.: spatialEco, R package version 1.3-6, 2021.

1194 Flanders Marine Institute: Boundies of Canda EEZ - mrgid 8493, Maritime Boundaries Geodatabase:
1195 Maritime Boundaries and Exclusive Economic Zones (200NM), version 11.,
1196 https://doi.org/10.14284/386, 2019.

1197 GEBCO: GEBCO Compilation Group - GEBCO_2022 Grid, The General Bathymetric Chart of the Oceans,
1198 https://doi.org/10.5285/e0f0bb80-ab44-2739-e053-6c86abc0289c, 2022.

1199 Gräler, B., Pebesma, E., and Heuvelink, G.: Spatio-Temporal Interpolation using gstat, The R Journal, 8,
1200 204–218, 2016.

1201 Graw, J. H., Wood, W. T., and Phrampus, B. J.: Predicting Global Marine Sediment Density Using the
1202 Random Forest Regressor Machine Learning Algorithm, Journal of Geophysical Research: Solid Earth,
1203 126, e2020JB020135, https://doi.org/10.1029/2020JB020135, 2021.

1204 Gregr, E. J., Haggarty, D. R., Davies, S. C., Fields, C., and Lessard, J.: Comprehensive marine substrate
1205 classification applied to Canada's Pacific shelf, PLOS ONE, 16, 1–28,
1206 https://doi.org/10.1371/journal.pone.0259156, 2021.

1207 Gudmundsson, L., Bremnes, J. B., Haugen, J. E., and Engen-Skaugen, T.: Technical Note: Downscaling
1208 RCM precipitation to the station scale using statistical transformations – a comparison of
1209 methods, Hydrology and Earth System Sciences, 16, 3383–3390, https://doi.org/10.5194/hess-16-3383-
1210 2012, 2012.

1211 Halpern, B. S., Frazier, M., Afflerbach, J., Lowndes, J. S., Micheli, F., O'Hara, C., Scarborough, C., and
1212 Selkoe, K. A.: Recent pace of change in human impact on the world's ocean, Scientific Reports, 9, 11609,
1213 https://doi.org/10.1038/s41598-019-47201-9, 2019.

1214 Hiddink, J. G., van de Velde, S. J., McConnaughey, R. A., De Borger, E., Tiano, J., Kaiser, M. J., Sweetman,
1215 A. K., and Sciberras, M.: Quantifying the carbon benefits of ending bottom trawling, Nature, 617, E1–E2,
1216 https://doi.org/10.1038/s41586-023-06014-7, 2023.

1217 Hijmans, R. J.: terra: Spatial Data Analysis, R package version 1.5-21, https://CRAN.R-
1218 project.org/package=terra, 2022.

1219 Hilborn, R. and Kaiser, M. J.: A path forward for analysing the impacts of marine protected areas, Nature,
1220 607, E1–E2, https://doi.org/10.1038/s41586-022-04775-1, 2022.

1221 Hoegh-Guldberg, O., Lovelock, C., Caldeira, K., Howard, J., Chopin, T., and Gaines, S.: The ocean as a
1222 solution to climate change: five opportunities for action, World Resources Institute, Washington, DC,
1223 2019.

1224 Hu, X., Myers, P. G., and Lu, Y.: Pacific Water Pathway in the Arctic Ocean and Beaufort Gyre in Two
1225 Simulations With Different Horizontal Resolutions, Journal of Geophysical Research: Oceans, 124, 6414–
1226 6432, https://doi.org/10.1029/2019JC015111, 2019.

1227 Hülse, D., Arndt, S., Wilson, J. D., Munhoven, G., and Ridgwell, A.: Understanding the causes and
1228 consequences of past marine carbon cycling variability through models, Earth-Science Reviews, 171,
1229 349–382, https://doi.org/10.1016/j.earscirev.2017.06.004, 2017.

1230 Ilich, A. R., Misiuk, B., Lecours, V., and Lecours, S. A.: MultiscaleDTM,
1231 https://doi.org/10.5281/zenodo.5548338, 2021.

1232 IPPC: 2019 Refinement to the 2006 IPCC Guidelines for National Greenhouse Gas Inventories, The
1233 International Plant Protection Convention, 2019.

1234 Jakobsson, M., Mayer, L. A., Bringensparr, C., Castro, C. F., Mohammad, R., Johnson, P., Ketter, T.,
1235 Accettella, D., Amblas, D., An, L., Arndt, J. E., Canals, M., Casamor, J. L., Chauché, N., Coakley, B.,
1236 Danielson, S., Demarte, M., Dickson, M.-L., Dorschel, B., Dowdeswell, J. A., Dreutter, S., Fremand, A. C.,
1237 Gallant, D., Hall, J. K., Hehemann, L., Hodnesdal, H., Hong, J., Ivaldi, R., Kane, E., Klaucke, I., Krawczyk, D.
1238 W., Kristoffersen, Y., Kuipers, B. R., Millan, R., Masetti, G., Morlighem, M., Noormets, R., Prescott, M. M.,
1239 Rebesco, M., Rignot, E., Semiletov, I., Tate, A. J., Travaglini, P., Velicogna, I., Weatherall, P., Weinrebe,
1240 W., Willis, J. K., Wood, M., Zarayskaya, Y., Zhang, T., Zimmermann, M., and Zinglersen, K. B.: The
1241 International Bathymetric Chart of the Arctic Ocean Version 4.0, Scientific Data, 7, 176,
1242 https://doi.org/10.1038/s41597-020-0520-9, 2020.

1243 Jenkins, C. J.: Summary of the onCALCULATION methods used in dbSEABED, in: Buczkowski, B.J., Reid,
1244 J.A., Jenkins, C.J., Reid, J.M., Williams, S.J., and Flocks, J.G. usSEABED: Gulf of Mexico and Caribbean
1245 (Puerto Rico and U.S. Virgin Islands) Offshore Surficial Sediment Data Release: U.S, United States
1246 Geological Survey, 2005.

1247 Keil, R.: Anthropogenic Forcing of Carbonate and Organic Carbon Preservation in Marine Sediments,
1248 Annual Review of Marine Science, 9, 151–172, https://doi.org/10.1146/annurev-marine-010816-
1249 060724, 2017.

1250 Kuhn, M.: caret: Classification and Regression Training, R package version 6.0-93, 2022.

1251 Kuhn, M. and Silge, J.: Tidy Modeling with R, O'Reilly Media, Inc, 2023.

1252 Kuhn, M. and Wickham, H.: Tidymodels: a collection of packages for modeling and machine learning
1253 using tidyverse principles, 2020.

1254 Kuzyk, Z. Z. A., Gobeil, C., Goñi, M. A., and Macdonald, R. W.: Early diagenesis and trace element
1255 accumulation in North American Arctic margin sediments, Geochimica et Cosmochimica Acta, 203, 175–
1256 200, https://doi.org/10.1016/j.gca.2016.12.015, 2017.

1257 LaRowe, D. E., Arndt, S., Bradley, J. A., Burwicz, E., Dale, A. W., and Amend, J. P.: Organic carbon and
1258 microbial activity in marine sediments on a global scale throughout the Quaternary, Geochimica et
1259 Cosmochimica Acta, 286, 227–247, https://doi.org/10.1016/j.gca.2020.07.017, 2020a.

1260 LaRowe, D. E., Arndt, S., Bradley, J. A., Estes, E. R., Hoarfrost, A., Lang, S. Q., Lloyd, K. G., Mahmoudi, N.,
1261 Orsi, W. D., Shah Walter, S. R., Steen, A. D., and Zhao, R.: The fate of organic carbon in marine sediments
1262 - New insights from recent data and analysis, Earth-Science Reviews, 204,
1263 https://doi.org/10.1016/j.earscirev.2020.103146, 2020b.

1264 Lee, T. R., Wood, W. T., and Phrampus, B. J.: A Machine Learning (kNN) Approach to Predicting Global
1265 Seafloor Total Organic Carbon, Global Biogeochemical Cycles, 33, 37–46,
1266 https://doi.org/10.1029/2018gb005992, 2019.

1267 Legge, O., Johnson, M., Hicks, N., Jickells, T., Diesing, M., Aldridge, J., Andrews, J., Artioli, Y., Bakker, D. C.
1268 E., Burrows, M. T., Carr, N., Cripps, G., Felgate, S. L., Fernand, L., Greenwood, N., Hartman, S., Kröger, S.,
1269 Lessin, G., Mahaffey, C., Mayor, D. J., Parker, R., Queirós, A. M., Shutler, J. D., Silva, T., Stahl, H., Tinker,
1270 J., Underwood, G. J. C., Van Der Molen, J., Wakelin, S., Weston, K., and Williamson, P.: Carbon on the
1271 Northwest European Shelf: Contemporary Budget and Future Influences, Frontiers in Marine Science, 7,
1272 Article 143, https://doi.org/10.3389/fmars.2020.00143, 2020.

1273 Ludwig, M., Moreno-Martinez, A., Hölzel, N., Pebesma, E., and Meyer, H.: Assessing and improving the
1274 transferability of current global spatial prediction models, Global Ecology and Biogeography, 32, 356–
1275 368, https://doi.org/10.1111/geb.13635, 2023.

1276 Luisetti, T., Turner, R. K., Andrews, J. E., Jickells, T. D., Kröger, S., Diesing, M., Paltriguera, L., Johnson, M.
1277 T., Parker, E. R., Bakker, D. C. E., and Weston, K.: Quantifying and valuing carbon flows and stores in
1278 coastal and shelf ecosystems in the UK, Ecosystem Services, 35, 67–76,
1279 https://doi.org/10.1016/j.ecoser.2018.10.013, 2019.

1280 Luisetti, T., Ferrini, S., Grilli, G., Jickells, T. D., Kennedy, H., Kröger, S., Lorenzoni, I., Milligan, B., van der
1281 Molen, J., Parker, R., Pryce, T., Turner, R. K., and Tyllianakis, E.: Climate action requires new accounting
1282 guidance and governance frameworks to manage carbon in shelf seas, Nature Communications, 11,
1283 4599, https://doi.org/10.1038/s41467-020-18242-w, 2020.

1284 Macreadie, P. I., Costa, M. D. P., Atwood, T. B., Friess, D. A., Kelleway, J. J., Kennedy, H., Lovelock, C. E.,
1285 Serrano, O., and Duarte, C. M.: Blue carbon as a natural climate solution, Nature Reviews Earth &
1286 Environment, 2, 826–839, https://doi.org/10.1038/s43017-021-00224-1, 2021.

1287 Madec, G., Delecluse, P., Imbard, M., and Lévy, C.: OPA 8.1Ocean General Circulation Model, Technical
1288 Report of LODYC/IPSL, Note 11, 1998.

1289 Martens, J., Romankevich, E., Semiletov, I., Wild, B., van Dongen, B., Vonk, J., Tesi, T., Shakhova, N.,
1290 Dudarev, O. V., Kosmach, D., Vetrov, A., Lobkovsky, L., Belyaev, N., Macdonald, R. W., Pieńkowski, A. J.,
1291 Eglinton, T. I., Haghipour, N., Dahle, S., Carroll, M. L., Åström, E. K. L., Grebmeier, J. M., Cooper, L. W.,
1292 Possnert, G., and Gustafsson, Ö.: CASCADE – The Circum-Arctic Sediment CArbon DatabasE, Earth
1293 System Science Data, 13, 2561–2572, https://doi.org/10.5194/essd-13-2561-2021, 2021.

1294 Martin, K. M., Wood, W. T., and Becker, J. J.: A global prediction of seafloor sediment porosity using
1295 machine learning, Geophysical Research Letters, 42, 10,640-10,646,
1296 https://doi.org/10.1002/2015GL065279, 2015.

1297 Masson, D. and Fine, I.: Modeling seasonal to interannual ocean variability of coastal British Columbia,
1298 Journal of Geophysical Research: Oceans, 117, https://doi.org/10.1029/2012JC008151, 2012.

1299 Maxwell, A. E. and Shobe, C. M.: Land-surface parameters for spatial predictive mapping and modeling,
1300 Earth-Science Reviews, 226, 103944, https://doi.org/10.1016/j.earscirev.2022.103944, 2022.

1301 Meyer, H. and Pebesma, E.: Machine learning-based global maps of ecological variables and the
1302 challenge of assessing them, Nature Communications, 13, 2208, https://doi.org/10.1038/s41467-022-
1303 29838-9, 2022.

1304   Meyer, H., Reudenbach, C., Wöllauer, S., and Nauss, T.: Importance of spatial predictor variable
1305   selection in machine learning applications – Moving from data reproduction to spatial prediction,
1306   Ecological Modelling, 411, 108815, https://doi.org/10.1016/j.ecolmodel.2019.108815, 2019.

1307   Meyer, H., Milà, C., and Ludwig, M.: CAST: "caret" Applications for Spatial-Temporal Models, R package
1308   version 0.7.1, 2023.

1309   Microsoft Corporation and Weston, S.: doParallel: Foreach Parallel Adaptor for the "parallel" Package, R
1310   package version 1.0.17, 2022.

1311   Middelburg, J. J.: Reviews and syntheses: to the bottom of carbon processing at the seafloor,
1312   Biogeosciences, 15, 413–427, https://doi.org/10.5194/bg-15-413-2018, 2018.

1313   Middelburg, J. J.: Marine Carbon Biogeochemistry: A Primer for Earth System Scientists, Springer, Cham,
1314   Switzerland, 118 pp., 2019.

1315   Mitchell, P. J., Aldridge, J., and Diesing, M.: Legacy Data: How Decades of Seabed Sampling Can Produce
1316   Robust Predictions and Versatile Products, Geosciences, 9,
1317   https://doi.org/10.3390/geosciences9040182, 2019.

1318   Molnar, C., Bischl, B., and Casalicchio, G.: iml: An R package for Interpretable Machine Learning, JOSS, 3,
1319   786, https://doi.org/10.21105/joss.00786, 2018.

1320   Nellemann, C., Corcoran, E., Duarte, C. M., Valdés, L., De Young, C., Fonseca, L., and Grimsditch, G.: Blue
1321   Carbon: A Rapid Response Assessment., United Nations Environment Programme, GRID-Arendal,
1322   Norway, 2009.

1323   NRCan: Lakes, Rivers and Glaciers in Canada - Hydrographic Features - Natural Resources Canada,
1324   Topographic Data of Canada - CanVec Series, 2019.

1325   NRCan: Canada west coast topo-bathymetric digital elevation model - Natural Resources
1326   Canada/Department of Fisheries and Oceans, Open Canada, 2021.

1327   NRCan: The Expedition Database (ED), Natural Resources Canada - Grain Size Data, https://ed.marine-
1328   geo.canada.ca/index_e.php, 2022.

1329   Pace, M. C., Bailey, D. M., Donnan, D. W., Narayanaswamy, B. E., Smith, H. J., Speirs, D. C., Turrell, W. R.,
1330   and Heath, M. R.: Modelling seabed sediment physical properties and organic matter content in the
1331   Firth of Clyde, Earth System Science Data, 13, 5847–5866, https://doi.org/10.5194/essd-13-5847-2021,
1332   2021.

1333   PANGAEA®: Data Publisher for Earth & Environmental Science, https://doi.org/10.1594/PANGAEA, 2022.

1334   Paradis, S., Nakajima, K., Van der Voort, T. S., Gies, H., Wildberger, A., Blattmann, T., Bröder, L., and
1335   Eglinton, T.: The Modern Ocean Sediment Archive and Inventory of Carbon (MOSAIC): version 2.0, Earth
1336   System Science Data, 2023, 4105–4125, https://doi.org/10.5194/essd-15-4105-2023, 2023.

1337   Pebesma, E.: Simple Features for R: Standardized Support for Spatial Vector Data, The R Journal, 10,
1338   439–446, https://doi.org/10.32614/RJ-2018-009, 2018.

1339   Pebesma, E.: stars: Spatiotemporal Arrays, Raster and Vector Data Cubes, R package version 0.6-0, 2022.

1340   Pedersen, T. L.: patchwork: The Composer of Plots, R package version 1.1.2, 2022.

1341   Peña, M. A., Fine, I., and Callendar, W.: Interannual variability in primary production and shelf-offshore
1342   transport of nutrients along the northeast Pacific Ocean margin, Deep Sea Research Part II: Topical
1343   Studies in Oceanography, 169–170, 104637, https://doi.org/10.1016/j.dsr2.2019.104637, 2019.

1344   Philibert, G., Todd, B. J., Campbell, D. C., King, E. L., Normandeau, A., Hayward, S. E., Patton, E. R., and
1345   Campbell, L.: Updated surficial geology compilation of the Scotian Shelf bioregion, offshore Nova Scotia
1346   and New Brunswick, Geological Survey of Canada - Open file, 8911, .zip file,
1347   https://doi.org/10.4095/330474, 2022.

1348   Posit Team: RStudio: Integrated Development Environment for R, Posit Software, PBC, Boston, MA,
1349   2022.

1350   Probst, P., Wright, M. N., and Boulesteix, A.-L.: Hyperparameters and tuning strategies for random
1351   forest, WIREs Data Mining and Knowledge Discovery, 9, e1301, https://doi.org/10.1002/widm.1301,
1352   2019.

1353   QGIS.org: QGIS Geographic Information System., QGIS Association, http://www.qgis.org, 2021.

1354   R Core Team: R: A language and environment for statistical computing., R Foundation for Statistical
1355   Computing, Vienna, Austria., https://www.R-project.org/, 2022.

1356   Raven, J.: Blue carbon: past, present and future, with emphasis on macroalgae, Biology Letters, 14,
1357   https://doi.org/10.1098/rsbl.2018.0336, 2018.

1358   Restreppo, G. A., Wood, W. T., Graw, J. H., and Phrampus, B. J.: A machine-learning derived model of
1359   seafloor sediment accumulation, Marine Geology, 440, 106577,
1360   https://doi.org/10.1016/j.margeo.2021.106577, 2021.

1361   Roy, M.-H. and Larocque, D.: Prediction intervals with random forests, Stat Methods Med Res, 29, 205–
1362   229, https://doi.org/10.1177/0962280219829885, 2020.

1363   Sala, E., Mayorga, J., Bradley, D., Cabral, R. B., Atwood, T. B., Auber, A., Cheung, W., Costello, C., Ferretti,
1364   F., Friedlander, A. M., Gaines, S. D., Garilao, C., Goodell, W., Halpern, B. S., Hinson, A., Kaschner, K.,
1365   Kesner-Reyes, K., Leprieur, F., McGowan, J., Morgan, L. E., Mouillot, D., Palacios-Abrantes, J.,
1366   Possingham, H. P., Rechberger, K. D., Worm, B., and Lubchenco, J.: Protecting the global ocean for
1367   biodiversity, food and climate, Nature, 592, 397–402, https://doi.org/10.1038/s41586-021-03371-z,
1368   2021.

1369   Seiter, K., Hensen, C., Schröter, J., and Zabel, M.: Organic carbon content in surface sediments—defining
1370   regional provinces, Deep Sea Research Part I: Oceanographic Research Papers, 51, 2001–2026,
1371   https://doi.org/10.1016/j.dsr.2004.06.014, 2004.

1372   Silge, J. and Mahoney, M.: spatialsample: Spatial Resampling Infrastructure, R package version 0.3.0,
1373   2023.

1374 Smeaton, C., Hunt, C. A., Turrell, W. R., and Austin, W. E. N.: Marine Sedimentary Carbon Stocks of the
1375 United Kingdom's Exclusive Economic Zone, Frontiers in Earth Science, 9, 50,
1376 https://doi.org/10.3389/feart.2021.593324, 2021.

1377 Snelgrove, P. V. R., Soetaert, K., Solan, M., Thrush, S., Wei, C. L., Danovaro, R., Fulweiler, R. W., Kitazato,
1378 H., Ingole, B., Norkko, A., Parkes, R. J., and Volkenborn, N.: Global Carbon Cycling on a Heterogeneous
1379 Seafloor, Trends in Ecology & Evolution, 33, 96–105, https://doi.org/10.1016/j.tree.2017.11.004, 2018.

1380 Soontiens, N. and Allen, S. E.: Modelling sensitivities to mixing and advection in a sill-basin estuarine
1381 system, Ocean Modelling, 112, 17–32, https://doi.org/10.1016/j.ocemod.2017.02.008, 2017.

1382 Soontiens, N., Allen, S. E., Latornell, D., Le Souëf, K., Machuca, I., Paquin, J.-P., Lu, Y., Thompson, K., and
1383 Korabel, V.: Storm Surges in the Strait of Georgia Simulated with a Regional Model, Atmosphere-Ocean,
1384 54, 1–21, https://doi.org/10.1080/07055900.2015.1108899, 2016.

1385 Sothe, C., Gonsamo, A., Arabian, J., Kurz, W. A., Finkelstein, S. A., and Snider, J.: Large Soil Carbon
1386 Storage in Terrestrial Ecosystems of Canada, Global Biogeochemical Cycles, 36, e2021GB007213,
1387 https://doi.org/10.1029/2021GB007213, 2022.

1388 Soulsby, R. L.: Simplified calculation of wave orbital velocities, Report TR 155 - HR Wallingford, 1, 2006.

1389 Stephens, D. and Diesing, M.: Towards Quantitative Spatial Models of Seabed Sediment Composition,
1390 PLOS ONE, 10, e0142502, https://doi.org/10.1371/journal.pone.0142502, 2015.

1391 Sumner, M.: tidync: A Tidy Approach to "NetCDF" Data Exploration and Extraction, R package version
1392 0.3.0, 2022.

1393 Turner, J. T.: Zooplankton fecal pellets, marine snow, phytodetritus and the ocean's biological pump,
1394 Prog. Oceanogr., 130, 205–248, https://doi.org/10.1016/j.pocean.2014.08.005, 2015.

1395 Van Rossum, G. and Drake, F. L.: Python 3 Reference Manual, CreateSpace, Scotts Valley, CA, 2009.

1396 VERRA: Methods for Monitoring of Carbon Stock Changes and Greenhouse Gas Emissions and Removals
1397 in Tidal Wetland Restoration and Conservation Project Activities (M-TW), VCS Module VMD0051,
1398 Sectoiral Scope 14, 1, 2020.

1399 Wager, S., Hastie, T., and Efron, B.: Confidence Intervals for Random Forests: The Jackknife and the
1400 Infinitesimal Jackknife, Journal of Machine Learning Research, 15, 1625–1651, 2014.

1401 Wickham, H., François, R., Henry, L., Müller, K., and Vaughan, D.: Welcome to the {tidyverse}, Journal of
1402 Open Source Software, 4, 1686, https://doi.org/10.21105/joss.01686, 2019.

1403 Wilson, R. J., Speirs, D. C., Sabatino, A., and Heath, M. R.: A synthetic map of the north-west European
1404 Shelf sedimentary environment for applications in marine science, Earth Syst. Sci. Data, 10, 109–130,
1405 https://doi.org/10.5194/essd-10-109-2018, 2018.

1406 Wood, S. N., Pya, N., and Säfken, B.: Smoothing Parameter and Model Selection for General Smooth
1407 Models, Journal of the American Statistical Association, 111, 1548–1563,
1408 https://doi.org/10.1080/01621459.2016.1180986, 2016.

1409   Wright, M., N. and Ziegler, A.: {ranger}: A Fast Implementation of Random Forests for High Dimensional
1410   Data in {C++} and {R}, Journal of Statistical Software, 77, 1–17, 2017.

1411   Wright, M. N., Ziegler, A., and König, I. R.: Do little interactions get lost in dark random forests?, BMC
1412   bioinformatics, 17, 1–10, 2016.

1413   Zhang, X., Chen, S., Xue, J., Wang, N., Xiao, Y., Chen, Q., Hong, Y., Zhou, Y., Teng, H., Hu, B., Zhuo, Z., Ji,
1414   W., Huang, Y., Gou, Y., Richer-de-Forges, A. C., Arrouays, D., and Shi, Z.: Improving model parsimony and
1415   accuracy by modified greedy feature selection in digital soil mapping, Geoderma, 432, 116383,
1416   https://doi.org/10.1016/j.geoderma.2023.116383, 2023.

1417