An observation-based evaluation and ranking of historical Earth System Model simulations for regional downscaling in the northwest North Atlantic Ocean

Arnaud Laurent[1], Katja Fennel[1], Angela Kuhn[2]

5   [1]Department of Oceanography, Dalhousie University, Halifax, Nova Scotia, Canada
    [2]Scripps Institution of Oceanography, UC San Diego, USA

Correspondence to: Arnaud Laurent (arnaud.laurent@dal.ca)

Abstract. Continental shelf regions in the ocean play an important role in the global cycling of carbon and nutrients but their responses to global change are understudied. Global Earth System Models (ESM), as essential tools for building understanding
10 of ocean biogeochemistry, are used extensively and routinely for projections of future climate states; however, their relatively coarse spatial resolution is likely not appropriate for accurately representing the complex patterns of circulation and elemental fluxes on the shelves along ocean margins. Here, we compared 29 ESMs used in the IPCC's Assessment Rounds (AR) 5 and 6 and a regional biogeochemical model for the northwest North Atlantic (NWA) shelf to assess their ability to reproduce observations of temperature, nitrate, and chlorophyll. The NWA region is biologically productive, influenced by the large-
15 scale Gulf Stream and Labrador Current systems, and particularly sensitive to climate change. Most ESMs compare relatively poorly to observed nitrate and chlorophyll and show differences with observed temperature due to spatial mismatches in their large-scale circulation. Model-simulated nitrate and chlorophyll compare better with available observations in AR6 than in AR5, but none of the models performs equally well for all 3 parameters. The ensemble means of all ESMs, and of the five best performing ESMs, strongly underestimate observed chlorophyll and nitrate. The regional model has a much higher spatial
20 resolution and reproduces the observations significantly better than any of the ESMs. It also simulates reasonably well vertically resolved observations from gliders and bi-monthly ship-based monitoring observations. A ranking of the ESMs suggests that the top 3 models are appropriate as boundary forcing for regional projections of future changes in the NWA region.

1 Introduction

25 Elemental fluxes along ocean margins, which are areas of complex physical and biogeochemical interactions, are important components of the global cycles of carbon (C) and nitrogen (N). For example, continental shelves host up to a third of oceanic primary production and over 40% of carbon burial in the ocean (Ducklow and McCallister, 2004; Muller-Karger, 2005; Walsh, 1991). They also are important sites of sediment denitrification leading to a net removal of fixed nitrogen (Fennel et al., 2006; Seitzinger and Giblin, 1996). Many shelf regions are thought to be a significant sink for atmospheric CO_2 (Cai et al., 2006;

30   Chen et al., 2013; Laruelle et al., 2018), including the eastern margin of North America (Fennel et al., 2019 and references therein), although there are significant discrepancies in available estimates. Despite their importance, the response of ocean margins to climate change is understudied relative to the open ocean.

Future projections of ocean biogeochemistry rely heavily on Earth System Models (ESMs). These are state-of-the-art, comprehensive representations of the major earth system components (including atmosphere, ocean, and land surface) and are

35   routinely used to perform climate scenario projections. The spatial resolution of the CMIP-class ESMs typically ranges from 0.5 to 2° and is too coarse to resolve coastal ocean dynamics and interactions between shelf and the open ocean (Anav et al., 2013; Bonan and Doney, 2018; Holt et al., 2017). This leads to uncertainty in future projections, not only for margin regions, and a global underestimation of the high primary productivity in coastal regions (Bopp et al., 2013; Schneider et al., 2008). Regional coupled circulation-biogeochemical models have been developed at much higher spatial resolution. These regional

40   models have been used to investigate biogeochemical processes along ocean margins (Fennel et al., 2006, 2013; Lachkar and Gruber, 2011; Peña et al., 2019; Siedlecki et al., 2015; Zhang et al., 2019) and project future states resulting from climate change (Gruber et al., 2012; Hermann et al., 2016; Holt et al., 2016; Laurent et al., 2018). The regional models allow for the temporal and spatial resolution necessary to resolve mesoscale processes and can be regionally calibrated (e.g., Kuhn and Fennel, 2019; Mattern and Edwards, 2017). However, the dynamics of a regional model is strongly determined by information

45   imposed along the model's open lateral boundaries, typically derived from a larger scale model, reanalysis product, or observation-based climatology. For future climate simulations, a regional model requires boundary information from future projections of large-scale models or ESMs.

The northwest North Atlantic (NWA), located at the confluence of the subtropical and subpolar gyres, is particularly challenging to global ocean circulation models and highly sensitive to climate-induced modifications of the large-scale

50   circulation, which are thought to be responsible for a multi-decadal deoxygenation trend in the region (Claret et al., 2018; Gilbert et al., 2010). While the CMIP models reasonably describe the large-scale climatological features of ocean physics in the NWA, the detailed current structure is poorly represented due to a mismatch in the location of the subtropical and subpolar gyres (Loder et al., 2015). The Gulf Stream usually extends too far north and the branch of the Labrador Current flowing southwest along the shelf edge tends to be missing (Lavoie et al., 2019; Loder et al., 2015). This leads to a warm bias in the

55   NWA, a common feature among coarse resolution ESMs (Saba et al., 2016). The absence of the shelf-break current significantly impacts cross-shelf exchange with much larger shelf water residence times in a high-resolution regional model (Rutherford and Fennel 2018) compared to estimates from a global model (Bourgeois et al. 2016). These discrepancies have been attributed to the coarse resolution of the global models (Loder et al., 2015, Saba et al. 2016, Rutherford and Fennel 2018, Lavoie et al., 2019).

60   Increased coastal model resolution can be achieved by downscaling large-scale or global models, the so-called parent models, to high-resolution regional models, the child models (see, e.g. Hermann et al. 2019, Holt et al. 2016, Laurent et al. 2018). For future projections, the obvious approach is to downscale ESMs. Since simulation of the fine-scale processes in the child model is strongly influenced by the parent model, it is important to assess the skill of ESMs in reproducing historical observations

65 prior to using them for downscaled future projections. Rickard et al. (2016) ranked ESMs based on their misfit with regional observations around New Zealand in order to discard models with significant errors and determine an ensemble of "best" models that can be used to study regional climate projections. Here, we take a similar approach.

Our objective is to assess the performance of a number of available ESMs in reproducing present conditions on the NWA shelf in contrast to a high-resolution regional model. More specifically, we compare 29 ESMs used in the two most recent IPCC Assessment Rounds (AR) as part of the Coupled Model Intercomparison Project 5 (CMIP5; Taylor et al., 2012) and its

70 currently ongoing successor CMIP6 (Eyring et al., 2016). We carry out a systematic and quantitative assessment and ranking by comparing the CMIP5 and CMIP6 models against observed surface temperature, chlorophyll, and nitrate and perform the same comparisons for a regional biogeochemical model. The latter is the Atlantic Canada Model (ACM, Brennan et al., 2016; Rutherford and Fennel, 2018) with biogeochemistry (Bianucci et al., 2016; Kuhn and Fennel, 2019) and is intended for regional downscaling of ESM simulations in order to generate high-resolution future projections. For all models, we present statistical

75 metrics based the mismatch of each model with climatological surface observations of temperature, nitrate, and chlorophyll and a ranking based on these metrics. The regional model is further evaluated against in-situ measurements, including high-resolution cross-shelf glider transects. The comparison provides sufficient confidence for only a few ESMs that their historical simulations can be used for downscaled future projections in the NWA. The use of ESM ensembles is not recommended. The regional model clearly outperformed all the global models.

2 Material and Methods

2.1 Models

2.1.1 Global models

The CMIP5 and CMIP6 framework provides state-of-the-art climate model datasets from the previous (AR5) and current (AR6) IPCC Assessment Rounds (Eyring et al., 2016; Taylor et al., 2012). Of all the ESMs, those that include ocean

85 biogeochemistry with monthly outputs of surface chlorophyll, nitrate, and temperature were included in our comparison. A total of 29 such ESMs were available (Table 1), 17 from CMIP5 (models 2–18, downloaded from the Earth System Grid Federation (ESGF) data repository at https://esgf-node.llnl.gov/search/cmip5/) and 12 from CMIP6 (models 19–30, downloaded from the ESGF data repository at https://esgf-node.llnl.gov/search/cmip6/). These models vary in their horizontal and vertical resolution and include a total of 13 different ocean biogeochemical models of varying levels of complexity (Table

90 1 and references therein).

We accessed the historical simulations which were forced by observed atmospheric composition and land cover changes over the period ~1850–2005 (CMIP5) and ~1850–2014 (CMIP6). Monthly, spatially resolved climatologies of surface chlorophyll, nitrate, and temperature were calculated over 30 years (1975–2005) from each ESM historical simulation.

2.1.2 Regional model

95   The ACM is a high-resolution, regional configuration of the Regional Ocean Modeling System (ROMS, version 3.5; Haidvogel et al., 2008) for the NWA, nested within the larger ocean-ice model of Urrego-Blanco and Sheng (2012), that includes the Gulf of Maine, Scotian Shelf and Grand Banks (Figure 1). The coupled physical-biogeochemical model has 30 vertical layers and an average horizontal resolution of 9.5 km on the shelf (Table 1). Detailed descriptions and physical model validation are presented in Brennan et al. (2016) and Rutherford and Fennel (2018). The biogeochemical model is based on Fennel et al.

100 (2006, 2008) but was expanded by splitting phytoplankton and zooplankton state variables into size-based functional groups, i.e. nano-micro-phytoplankton and micro-meso-zooplankton. The model was also modified by including temperature-dependent biological rates for nutrient uptake, phytoplankton and zooplankton mortality, grazing and zooplankton egestion and excretion (see supporting text). The model has 10 state variables: nitrate, ammonium, and two size classes each for phytoplankton, chlorophyll, zooplankton and detritus (Figure 2). This ecosystem structure is of intermediate complexity similar

105 to the model of Aumont et al. (2015), which is used in 6 of the ESMs included in our study. Model parameters were optimized by Kuhn (2017) and are listed in supporting Table S1. The model description and equations are available in the Supporting Information.

Initial and open boundary conditions for nitrate (NO_3) were defined from a monthly climatology (Kuhn, 2017) based on in-situ observations and the World Ocean Atlas 2009 (Garcia et al., 2010). Other biological variables were set to 0.1 mmol N m^{-3}

110 with a phytoplankton-to-chlorophyll ratio of 0.76 mmol N (mg Chl)$^{-1}$ (Bianucci et al., 2016). The model was initialized on January 1, 1999 and run through December 31, 2014. The first year was considered spin up. Monthly climatologies of surface chlorophyll, nitrate, and temperature were calculated for comparison with the ESMs.

2.1.3 Model resolution

The 30 models differ dramatically in their horizontal resolution and do not evenly cover the 3 regions of interest (Figure 3,

115 Table 1). The regional ACM has a much higher resolution than any of the ESMs with about 16 times more horizontal grid cells than the highest resolution ESM and almost 300 times more than the lowest resolution ESM. Among the ESMs the highest resolution is achieved by models 16 and 28, which share the same grid. These two have more than twice the number of horizontal grid cells than the next highest resolution models (3, 18, 20–21). The lowest resolution ESMs are models 3 and 12–14 with only 26 horizontal grid cells within the NWA shelf resulting in a coarse representation, particularly in the SS region.

120 The median number of grid cells in the NWA shelf region is 72 and 102 for the CMIP5 and CMIP6 models, respectively, compared to 6875 in the ACM.

2.2 Observations

Three types of observations were used in the model intercomparison: 1) satellite surface chlorophyll observations from the Sea-viewing Wide Field-of-view Sensor (SeaWiFS) as 8-day averaged maps at 1/12˚ resolution (1999–2010), 2) surface nitrate

125 from the World Ocean Atlas 2009 (WOA; Garcia et al., 2010) at 1˚ resolution, and 3) surface temperature from the Operational
SST and Sea Ice Analysis (OSTIA) system (Donlon et al., 2012) at 1/20˚ resolution (2006–2016). Monthly climatologies were
calculated for each of these.

In addition, the regional model was validated using high-resolution in-situ observations from the Atlantic Zone Monitoring
Program (AZMP, http://www.meds-sdmm.dfo-mpo.gc.ca/isdm-gdsi/azmp-pmza/index-eng.html) and glider transects along
130 the Halifax Line between 2011 and 2016 (Ross et al., 2017). Transect data from AZMP and gliders were seasonally averaged
along the Halifax Line (Figure 1).

2.3 Comparison metrics

For comparison with the observations, each model was mapped onto the SeaWiFS, WOA and OSTIA grids. Since some areas,
such as the nearshore and the Bay of Fundy, are covered by only a few models, grid cells that are active in less than 85% of
135 all models were excluded from the analysis to avoid biases. In the low-resolution WOA climatology, the months November to
January were excluded because poor data availability in these months resulted in unrealistic patterns.

Three zones were defined for a high-level comparison with the observations: the Gulf of Maine (GoM), Scotian Shelf (SS),
and Grand Banks (GB) (Figure 1). Subsequently, the term NWA shelf refers to the region covered by all 3 zones (GoM, SS
and GB).

140 Following the method of Rickard et al. (2016), a score S is calculated for each model variable, v (i.e., surface temperature,
chlorophyll, and nitrate), for each month, t, in the climatology as the sum of the centered Root Mean Square Difference
(RMSD) and bias between the observations (x) and the model (y), such that:

$$S(t,v) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left((x_i(t,v) - \bar{x}(t,v)) - (y_i(t,v) - \bar{y}(t,v))\right)^2} + \frac{1}{n}\left|\sum_{i=1}^{n}(x_i(t,v) - y_i(t,v))\right|$$

where the index i refers to a grid cell and n is the total number of grid cells within the NWA shelf. The lower the score the
145 better the match between model and observations. Annual mean scores $\bar{S}(v)$ were calculated for each model variable by
averaging over t. For each variable, the models were ranked based on their annual mean score. The overall rank was determined
by ranking models by the averages of their ranks for surface temperature, chlorophyll, and nitrate. For models with equal
averages the ranking was determined by the average of chlorophyll and nitrate ranks.

To facilitate the comparison with observations, the ESMs were grouped into CMIP5 and CMIP6 and the ensemble means of
150 all models and of the 5 highest ranked models were calculated for each group.

3 Results

Models and model ensembles are first compared with observations to assess their ability to reproduce the annual cycles of
surface temperature, chlorophyll and nitrate in the NWA region. Error statistics are then analyzed to understand how the models

deviate from each observed variable and subsequently used to calculate the scores and then rank the models. Finally, additional,
155   high-resolution comparisons between models and observations are presented to further assess the regional model's
performance.

3.1 Model-data comparisons

First, we compare the spatially averaged climatological surface temperature (Figure 4Figure 5a-c), chlorophyll (Figure 4Figure
5d-f) and nitrate (Figure 4Figure 5g-i) in our 3 regions of interest. The ESMs reasonably reproduce the annual cycle of surface
160   temperature, but the annual cycles of chlorophyll and nitrate are not simulated well in any of them (see supporting Figures S1-
S5) and the range of simulated biological properties is large.

Temperature is relatively consistent between model ensembles (Figure 4a–c), but with large variability between models (Figure
5a–c). An annual, positive bias occurs in the GM (bias = +2.28°C, Figure 4a), whereas winter temperatures are overestimated
on the SS and GB (bias = +2.32 and +1.26°C respectively, Figure 4a–c) and underestimated in summer on GB (bias = +2.32
165   and –1.93°C, Figure 4f).

For surface chlorophyll, there is a large discrepancy between the model ensembles and observations (Figure 4d–f). Inter-model
differences are largest for the time of maxima and magnitude of the spring and fall blooms (Figure 5d–f, supporting Figures
S1- S5). Standard deviations for the magnitude of the spring bloom are large among ESMs in the 3 zones (SD=0.6, 0.81 and
0.83 mg m^{-3} in GoM, SS and GB, respectively). The maxima of the spring bloom also varies significantly in time among the
170   models, with a standard deviation among ESMs for the time of maxima of the bloom of about 1.5 months (SD=1.15, 1.59 and
1.62 months in GoM, SS and GB, respectively). Most models in the CMIP5 group do not simulate a fall bloom, hence none is
present in the ESM ensemble mean, but rather a fall/winter increase in chlorophyll concentrations. Among the CMIP6 group,
only models 23–25 generate a fall bloom (see supporting Figures S4–S5). Overall, the ESMs underestimate annual surface
chlorophyll concentrations (bias = –0.93, –0.49 and –0.29 mg m^{-3} for GM, SS and GB, respectively, Figure 4a-c). The
175   chlorophyll bias is about 20% smaller in the CMIP6 group compared to CMIP5.

There are also large discrepancies between the model ensembles and observations for nitrate (Figure 4g–i), particularly in the
CMIP5 group. The variability in nitrate concentrations among the ESMs is also large (SD = 2.80 mmol m^{-3}) but smaller by
29% in the CMIP6 group. Most of the models reproduce the seasonal variability of surface nitrate (Figure 5g–i, supporting
Figures S1–S5); however, the CMIP5 models tend to underestimate fall-winter concentrations (winter bias = –1.28 mmol m^{-3}),
180   whereas the CMIP6 model group performs better but with some mismatches in the timing of the seasonal changes (spring,
fall). A few models markedly overestimate surface nitrate concentrations in the NWA shelf regions (see supporting Figures
S1, S3–5), including within the CMIP6 group. Supporting Figures S6–S9 provide an illustration of the model variability for
chlorophyll and nitrate in March (Figures S6 and S7) and October (Figures S8 and S9), i.e. around the time of the spring and
fall blooms respectively.
185   The regional ACM well reproduces the annual cycle of surface temperature (Figure 4a-c), chlorophyll (Figure 4d-f) and nitrate
(Figure 4g-i) in the three regions. The model correctly simulates the overall magnitude of temperature and chlorophyll biomass,

the timing of the maxima of spring and fall blooms and the latitudinal gradients in temperature, chlorophyll and nitrate, although the magnitude of the spring bloom in the GM and GB regions is underestimated.

3.2 Model statistics

190   Error statistics, i.e. RMSD and bias, are now analyzed and used to calculate the model scores. The distribution and relationships between scores are explored and then the ranks calculated.

The RMSD between the spatially averaged climatological observations and models are not consistent between variables, as indicated by the increasing temperature RMSD in Figure 6. However, temperature and chlorophyll RMSD are correlated ($r =$ 0.51, $p = 0.0043$). For temperature, models 3, 20–21, and 24–25 have the largest discrepancy with observations and some

195   clearly represent better the annual cycle than others. For chlorophyll, the largest discrepancies with observations are in models 4, 8 14 and 19–21, but overall chlorophyll RMSD are relatively large and homogeneous, except for a few models that have lower RMSD (e.g. models 22–23). Interestingly, the magnitude of the spring bloom in model 18 (CMIP5 group) is somewhat close to the observations. However, the time shift of the bloom (May–June) results in a poor agreement with observations. The mismatch between observed and simulated nitrate is much higher for models 5, 7, 18 and 29 and some models are much better

200   at representing the observed annual cycle (Figure 6). The models with lowest RMSD for all 3 parameters are models 22–23 (CMIP6 group). The RMSDs of the ACM are about a third of the average RMSD of the ESMs for both chlorophyll (ESM RMSDs are ×2.0–4.1 that of the ACM) and nitrate (×1.4–11.4) and a quarter for temperature (×1.1–10.4).

Model scores (see Sect. 2.3) represent the spatial and temporal mismatch within the NWA shelf region (Figure 7). In general, the scores provide similar results as the RMSDs in Figure 6, although groups tend to emerge from the score calculation. As

205   observed previously in Figure 6, the scores of ESMs have a much larger range of variability for temperature (1.5–7.8) and nitrate (1.4–13.2) than for chlorophyll (0.81–1.42) due to the large mismatch observed with a few models (Figure 7, supporting Figures S1–S5). For temperature, 4 of the 6 poorest (largest) scores (> 4.5) are in the CMIP6 group. They all markedly overestimate temperature, especially in the GM (see supporting Figures S1, S4–S5). The range of variability in chlorophyll scores did not reduce from CMIP5 to CMIP6 and given the improvement of a few CMIP6 models (i.e. 22 and 23), the range

210   is larger in the CMIP6 group (0.8–1.4, Figure 7, right panel) than in the CMIP5 group (1–1.4, Figure 7, left panel). With the exception of model 29, which has a very poor (high) score for nitrate, the range of variability in nitrate is reduced in the CMIP6 group. In total, 5 models (3, 5, 7, 18, 29) have very poor scores for nitrate (> 4) strongly overestimating surface nitrate, except for model 3 in the Gulf of Maine (see supporting Figure S1). The remaining models have more homogeneous nitrate scores (Figure 7) with the best (lowest) scores in models 25, 24, 9 and 6 (Table 2). Models that underestimate nitrate (2, 8, 14 and 19,

215   see supporting Figures S1–S4) have a better score because they match the low nitrate observations in late spring–summer (Table 2). Overall, ACM has the best scores, $\bar{S}(v)$, for temperature (1.14), chlorophyll (0.64) and nitrate (1.27).

Among the 3 variables, and including the regional model, we found a correlation between the scores of chlorophyll and temperature ($r = 0.53$, $p = 0.0025$), but not between nitrate and chlorophyll ($r = 0.03$, $p = 0.88$) or nitrate and temperature ($r = 0.06$, $p = 0.74$). As can be seen in Figure 6, the ESMs with a poor representation of nitrate are not necessarily performing

220   poorly with respect to chlorophyll or temperature. Model 7 for instance has the poorest score for nitrate and a relatively poor score for temperature but the best score of the CMIP5 group for chlorophyll (Figure 7, left panel). In fact, only models 3 and 18 have poor scores for all variables. Similarly, models 24 and 25 have the best scores for chlorophyll among the ESMs but are among the worst for temperature. On average, models have worse scores in the GM (3.97, 1.73, 3.15) than on the SS (3.35, 0.94, 2.22) and GB (2.53, 0.72, 2.46) for temperature, chlorophyll and nitrate, respectively.

225   Overall, 4 groups emerge on the chlorophyll-nitrate space in Figure 7. This grouping is somewhat arbitrary but follows the general ranking presented in Figure 8, with a few exceptions. Group A includes the 14 best models (6 CMIP5 and 8 CMIP6) except for model 9 and 30 whose ranking is degraded due to poor representation of temperature. Group B includes the 4 intermediate-score models (15, 16, 17, 2). Group C includes the 8 models with poor chlorophyll scores (5 CMIP5 and 3 CMIP6) and Group D the 5 models with poor nitrate scores (4 CMIP5 and 1 CMIP6). Most of the models with poor scores for
230   temperature are included in Group C, i.e. with the poor chlorophyll scores.

The overall model ranking (average of chlorophyll, nitrate and temperature ranks) indicates the gap between ACM and ESMs, as well as within ESMs (Figure 8). As expected, ACM ranks first, following the best scores for both chlorophyll and nitrate. The gap between ACM and model 22 (the best overall ESM) indicates that none of the ESM performs best for both chlorophyll and nitrate. This is also shown by the large range in individual ranks (dark grey lines in Figure 8) in most models. Group A
235   includes the 5 best ranking models, all from CMIP6 (22, 28, 25, 24, 23, respectively). The most consistent in term of chlorophyll and nitrate ranking is model 28, the other ones having a relatively large spread. The best ranked CMIP5 models are 10 and 13. On the other side of the spectrum models 20, 3, 21 and 18 (Groups C and D) have the poorest ranks because of their consistently poor scores for chlorophyll and nitrate. Despite their poor performance with respect to nitrate, models 7 and 29 are ranked within the mid-range of the ESMs because they are among the best ESMs with respect to chlorophyll (rank 4
240   and 8, respectively).

3.3 Additional model-data comparisons for regional ACM

While the resolution of the ESMs does not allow for a comparison at smaller spatial scales, we further compare the regional ACM to cross-shelf transects and station observations (Figure 9) along the Halifax Line (see Figure 1). The ACM reproduces the seasonal variation and the vertical gradient in chlorophyll and nitrate along the transect (Figure 9), although the simulated
245   distributions are smoother than the glider observations. The summer subsurface chlorophyll maximum is located at the appropriate depth (28 m simulated versus 32 m observed, on average). The ACM somewhat underestimates the depth of the nitracline in the offshore waters (34 m versus 43 m, $x > 150$ km) and overestimates surface nitrate in spring and fall, as seen in Figure 4.

Station 2, which is located nearshore on the Halifax Line (see Figure 1), provides additional, vertically resolved information
250   with high temporal resolution that is useful for model validation (Figure 10). At this location, the ACM reproduces the annual cycle of chlorophyll and nitrate. Surface and subsurface nitrate and chlorophyll are qualitatively reproduced in all seasons

except during the spring bloom, which is more pronounced and reaches deeper in the observations, although the magnitude and vertical distribution of chlorophyll concentration agree well with the glider observations at this time.

A quantitative, point-to-point comparison of the ACM with the time series and glider observations along the Halifax Line (Figure 9) and at Station 2 (Figure 10) is provided in Table 3. The comparison indicates relatively high correlations between the ACM and time series of chlorophyll (0.68–0.78) and nitrate (0.83–0.92) along the Halifax Line as well as glider measurements of chlorophyll (0.85–0.94) for all seasons. Correlations are high as well at Station 2 for nitrate time series and glider measurements of chlorophyll. The largest discrepancies with observations are found with the time series of chlorophyll in spring. These results, indicate an overall good skill of the model to reproduce the seasonal, vertically resolved observations on the Scotian Shelf.

4 Discussion

4.1 Overall model performance

There are significant discrepancies with observations and a large variability among ESMs in the representation of surface temperature, chlorophyll and nitrate in the NWA shelf (Table 2, Figure 6 and supporting Figures S1–S5). A warm bias resulting from a mismatch in the location of the Gulf Stream was present in most models, in line with the previous results of Loder et al. (2015) and Saba et al. (2016). Chlorophyll concentration was also systematically underestimated. The spring and fall blooms, which are characteristic annual features of the NWA region (Greenan et al., 2004, 2008) were absent in some and most models, respectively. The correlation between temperature and chlorophyll scores indicated that errors in surface chlorophyll concentration were likely driven by the misrepresentation of the general circulation and, more generally, of ocean physics.

Following Rickard et al. (2016), who used a similar ranking procedure, the 29 ESMs can be divided into an inner and an outer model ensemble. The outer ensemble includes 17 models that clearly misrepresent surface conditions in the NWA shelf (models 2–5, 7–8, 11, 14–16, 18–21, 24–25 and 29) and were selected as follows. The 8 models with lowest ranks (2–4, 8, 18–21) were included because they consistently misrepresent surface fields on the NWA shelf. Five of those were different generations (CMIP5 and CMIP6) of the same model, i.e. CanESM (2, 19) and CESM (3, 20–21). Their large scores imply that CanESM and CESM have fundamental issues with representing biogeochemistry in the NWA. Models 15–16 and 24–25 were also included in the outer ensemble because of their misrepresentation of surface nitrate and temperature, respectively. Since nitrate scores neither correlate with chlorophyll nor temperature, the mismatch with nitrate observations is likely related to intrinsic biogeochemical model behaviour rather than to a mismatch in circulation. Models with persistent positive or negative biases in surface nitrate (4–5, 7–8, 11, 14, 19 and 29, Figures S1–S5) were selected because they misrepresent the seasonal nitrate dynamics and therefore the other biogeochemical variables driven by nitrate are questionable.

The inner ensemble includes 12 models (6, 9–10, 12–13, 17, 22–23, 26–28, 30). Can those be used as a multi-model (optimal) ensemble to characterize the future state of the NWA shelf region? Unfortunately, we found that an ensemble mean of these

models, and even of the best five models, poorly represents historical surface fields due to the large variability within the
285 ensemble (Figure 5) and the biases in the ensemble surface temperature and chlorophyll concentration (Figure 4).

The regional model clearly outperformed the ESMs in our assessment, with a consistent representation of the surface and subsurface fields in all shelf areas. The high spatial resolution of the regional model also allowed for a fine scale model validation that was not possible for the ESMs. The complementary glider transects and time series stations provide a high-resolution dataset of in-situ chlorophyll and nitrate concentrations and shows that the regional model resolves seasonal and
290 vertical variation in chlorophyll and nitrate on the Scotian Shelf, something that none of the ESMs were able to reproduce.

4.2 Impact of spatial resolution

In general, the coarse horizontal resolution of the ESMs affects the representation of the NWA region in comparison to the regional model, particularly on the relatively narrow Scotian Shelf. The poor representation of coastal areas is a known limitation of global models (Holt et al., 2017) and results in a global underestimation of primary productivity in these regions
295 (Bopp et al., 2013; Schneider et al., 2008).

There is no correlation between grid resolution and ESM rank (Figure 11) despite the fact that the best ranked ESM (MPI-ESM1-2-HR) has also the highest resolution (Table 1Table 2). This result shows that higher grid resolution, as called for by Lavoie et al. (2013) for the NWA and by McKiver et al. (2015) for the global ocean, is not a guarantee for improved model performance. In fact, some very coarse resolution models from the CMIP5 group were ranked as well or better than the other
300 models and models with the second highest resolution (3, 18, 20–21) had all low ranks. The improved ranks at constant (e.g. models 22, 24, 25, 28) and even lower (model 29) ocean grid resolution in the CMIP6 group (Table 2, Figure 12) was also an indication that the discrepancies with observations, and the improvement in the CMIP6 models (see below), were not associated with the ocean grid resolution but rather resulted from the physical and biogeochemical setup of the models. Another hint at the lack of relationship between resolution and model rank is the similar ranking of the two MPI models in the CMIP5 group,
305 MPI-ESM-LR and MPI-ESM-MR, despite an important difference in model grid resolution (Figure 8). Much higher resolution will be necessary to refine the projections in coastal areas (e.g., Holt et al. (2017), Saba et al. (2016)), which is not currently computationally feasible in ESMs (Holt et al., 2009, 2017).

4.3 Impact of biogeochemical model structure

Although model performance is likely influenced by the biogeochemical model structure, we did not find a clear relationship
310 between biogeochemical model and performance. While the inner and outer ensembles share only 4 biogeochemical models (PISCES, HAMOCC, TOPAZ2, NOBM) out of 13, there was no indication of consistently better performance for the biogeochemical models in the inner ensemble. For example, models using similar ocean biogeochemistry (e.g., PISCES: 5, 12–14, 22, 26, and HAMOCC: 15–16, 18, 28–29) had very different ranks, with no obvious relationship between overall model rank and the ocean biogeochemical model component. Moreover, 4 biogeochemical models were represented in the 5 best
315 ranked ESMs, similar to previous findings by Rickard et al. (2016).

4.4 Improvement from CMIP5 to CMIP6

Model performance improved in the new CMIP generation, but not uniformly across models and variables. The 4 best ranked ESMs were from the CMIP6 group, although the average rank was not very different between the two groups, i.e. $\bar{R} = 17.4$ and 14.0 for CMIP5 and CMIP6, respectively (Figure 8, Table 2). The change in performance between the two generations of
320 models can be assessed by evaluating the subset of models that are available for CMIP5 and CMIP6. There are nine such models (Figure 12). All CMIP6 models have improved overall ranks, indicating better performance (Figure 12). The overall improvement was large only for models that had average to low ranks in the CMIP5 group (ranks 15–22, x-axis in Figure 12). Temperature did not improve except for GFDL-ESM2M and degraded in some cases. The change in ranking is therefore mainly associated with better surface fields for chlorophyll and nitrate. This is particularly the case for model pairs 3, 5, 6 and
325 8, which ranked much better for chlorophyll (+8.2) and nitrate (+12.7) in the CMIP6 group (Figure 12). The chlorophyll rank in model pair 4 improved significantly (+18) but this improvement was counteracted by degraded temperature and nitrate ranks. The lack of improvement in surface temperature indicates that the temperature bias detected in the CMIP5 group was not solved in CMIP6, as seen in Figure 4.

We can only speculate about the source of improvement in the CMIP6 models. Kwiatkowski et al. (2020) recently showed
330 that projected surface temperature, nitrate and net primary production differ significantly in CMIP5 and CMIP6 model ensembles. Higher climate sensitivity in CMIP6 models partly explain this difference but the source of change in primary production was not resolved. In the historical simulations, better surface chlorophyll and nitrate fields in CNRM-ESM2-1 may be associated with the transition from a climate model with ocean biogeochemistry to a fully coupled ESM, even though such transition may degrade historical simulations due to the replacement of observations by prognostic schemes that are poorly
335 constrained (Séférian et al., 2019). Updated land and ocean biogeochemistry may have improved the representation of surface chlorophyll and nitrate in MPI-ESM1-2-HR (Müller et al., 2018), whereas the improvement in surface temperature and nitrate fields from GFDL-ESM2M to GFDL-ESM4 seem to be associated with the physical ocean component of the model, given that GFDL-ESM2G already performed well in the CMIP5 group. Danabasoglu et al. (2020) found a significant improvement for CESM2 at the global scale but a poor representation of the Gulf Stream–North Atlantic Current system, resulting in a large
340 surface temperature bias. This is in line with our assessment for the NWA shelf where both physical and biological parameters had poor scores and the model was not found appropriate for shelf studies in the NWA.

4.5 Other coastal regions

Our results may also apply for other coastal regions, given the poor representation of coastal areas in ESMs, but the details are probably region specific. Discrepancies with observations in the NWA are partly driven by poor representation of large-scale
345 circulation features such as the Gulf Stream and Labrador Current in most of the models. The representation of large-scale currents may improve (or worsen) in other regions, resulting in a different ranking there. For example, Rickard et al. (2016) found a different model selection in the inner model ensemble around New Zealand. Seven (out of 11) of their inner ensemble

models (models 2–5, 7–8, 14) are not included in our inner ensemble. Model 3, perhaps the best model in their assessment, ranked 29 out of 30 in the NWA shelf region (Figure 8, supporting Figure S1). The representation of the dynamic NWA
350 circulation is a known issue in ESMs and further regional comparisons will be necessary to assess if our results are representative for the global coastal ocean.

5 Conclusions

We evaluated the CMIP5 and CMIP6 ESMs with biogeochemistry for the NWA shelf. Arguably, only 3 models (CNRM-ESM2-1, GFDL-ESM4 and MPI-ESM1-2-HR), all from CMIP6, provided sufficient confidence in their historical simulations.
355 We caution against using model ensembles, either directly or in downscaling future projections for the NWA shelf. The regional model (ACM) clearly outperformed the global models and is a good candidate for downscaled projections in combination with one of the top ranked ESMs. Further refinement in the ACM should focus on the mechanisms that determine the magnitude of the spring bloom.

Similar comparisons should be carried out in coastal areas before using CMIP model projections. While it is not clear how the
360 presented model ranking will hold in other regions, it is highly likely that some models do not perform well in coastal areas generally and should not be used for regional investigations.

Given the lack of a direct relationship between model skill and horizontal resolution, it is unlikely that feasible grid refinement will significantly improve model performance in the NWA region. The improvement in scores from CMIP5 to CMIP6 shows that refining ocean biogeochemical components can improve the model performance.

365 *Code and data availability.* The ROMS code and the observations are available from the links referenced in the manuscript.

Supplement link. The supplement related to this article is available on-line at:

Author contribution. AL and KF conceived the study. AL and AK set up the ACM model. AL conducted the analyses. AL wrote the manuscript with input from KF

Competing interests. The authors declare that they have no conflict of interest

370 *Acknowledgements.* The ACM was run on Compute Canada resources under the resource allocation project qqh-593-ac.

Financial statement. We acknowledge funding from the Canada First Research Excellence Fund, through the Ocean Frontier Institute, the MEOPAR Network of Centres of Excellence through the Prediction Core, and an NSERC Discovery Grant held by KF.

References

375 Adachi, Y., Yukimoto, S., Deushi, M., Obata, A., Nakano, H., Tanaka, T. Y., Hosaka, M., Sakami, T., Yoshimura, H., Hirabara, M., Shindo, E., Tsujino, H., Mizuta, R., Yabu, S., Koshiro, T., Ose, T. and Kitoh, A.: Basic performance of a new earth system model of the Meteorological Research Institute (MRI-ESM1), Pap. Meteorol. Geophys., 64, 1–19, doi:10.2467/mripapers.64.1,

2013.

Anav, A., Friedlingstein, P., Kidston, M., Bopp, L., Ciais, P., Cox, P., Jones, C., Jung, M., Myneni, R. and Zhu, Z.: Evaluating
380   the Land and Ocean Components of the Global Carbon Cycle in the CMIP5 Earth System Models, J. Clim., 26(18), 6801–
6843, doi:10.1175/JCLI-D-12-00417.1, 2013.

Arora, V. K., Scinocca, J. F., Boer, G. J., Christian, J. R., Denman, K. L., Flato, G. M., Kharin, V. V., Lee, W. G. and
Merryfield, W. J.: Carbon emission limits required to satisfy future representative concentration pathways of greenhouse gases,
Geophys. Res. Lett., 38(5), L05805, doi:10.1029/2010GL046270, 2011.

385   Aumont, O. and Bopp, L.: Globalizing results from ocean in situ iron fertilization studies, Global Biogeochem. Cycles, 20(2),
GB2017, doi:10.1029/2005GB002591, 2006.

Aumont, O., Ethé, C., Tagliabue, A., Bopp, L. and Gehlen, M.: PISCES-v2: An ocean biogeochemical model for carbon and
ecosystem studies, Geosci. Model Dev., 8(8), 2465–2513, doi:10.5194/gmd-8-2465-2015, 2015.

Bianucci, L., Fennel, K., Chabot, D., Shackell, N. and Lavoie, D.: Ocean biogeochemical models as management tools: a case
390   study for Atlantic wolffish and declining oxygen, ICES J. Mar. Sci. J. du Cons., 73(2), 263–274, doi:10.1093/icesjms/fsv220,
2016.

Bonan, G. B. and Doney, S. C.: Climate, ecosystems, and planetary futures: The challenge to predict life in Earth system
models, Science (80-.)., 359(6375), eaam8328, doi:10.1126/science.aam8328, 2018.

Bopp, L., Resplandy, L., Orr, J. C., Doney, S. C., Dunne, J. P., Gehlen, M., Halloran, P., Heinze, C., Ilyina, T., Séférian, R.,
395   Tjiputra, J. and Vichi, M.: Multiple stressors of ocean ecosystems in the 21st century: projections with CMIP5 models,
Biogeosciences, 10(10), 6225–6245, doi:10.5194/bg-10-6225-2013, 2013.

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., Bekki, S., Bonnet, R., Bony, S., Bopp,
L., Braconnot, P., Brockmann, P., Cadule, P., Caubel, A., Cheruy, F., Codron, F., Cozic, A., Cugnet, D., D'Andrea, F., Davini,
P., Lavergne, C., Denvil, S., Deshayes, J., Devilliers, M., Ducharne, A., Dufresne, J., Dupont, E., Éthé, C., Fairhead, L., Falletti,
400   L., Flavoni, S., Foujols, M., Gardoll, S., Gastineau, G., Ghattas, J., Grandpeix, J., Guenet, B., Guez, L., Guilyardi, É.,
Guimberteau, M., Hauglustaine, D., Hourdin, F., Idelkadi, A., Joussaume, S., Kageyama, M., Khodri, M., Krinner, G., Lebas,
N., Levavasseur, G., Lévy, C., Li, L., Lott, F., Lurton, T., Luyssaert, S., Madec, G., Madeleine, J., Maignan, F., Marchand,
M., Marti, O., Mellul, L., Meurdesoif, Y., Mignot, J., Musat, I., Ottlé, C., Peylin, P., Planton, Y., Polcher, J., Rio, C., Rochetin,
N., Rousset, C., Sepulchre, P., Sima, A., Swingedouw, D., Thiéblemont, R., Traore, A. K., Vancoppenolle, M., Vial, J., Vialard,
405   J., Viovy, N. and Vuichard, N.: Presentation and evaluation of the IPSL-CM6A-LR climate model, J. Adv. Model. Earth Syst.,
doi:10.1029/2019MS002010, 2020.

Bourgeois, T., Orr, J. C., Resplandy, L., Terhaar, J., Ethé, C., Gehlen, M. and Bopp, L.: Coastal-ocean uptake of anthropogenic
carbon, Biogeosciences, 13(14), 4167–4185, doi:10.5194/bg-13-4167-2016, 2016.

Brennan, C. E., Bianucci, L. and Fennel, K.: Sensitivity of Northwest North Atlantic shelf circulation to surface and boundary
410   forcing: a regional model assessment, Atmosphere-Ocean, 54(3), 230–247, doi:10.1080/07055900.2016.1147416, 2016.

Cai, W.-J., Dai, M. and Wang, Y.: Air-sea exchange of carbon dioxide in ocean margins: A province-based synthesis, Geophys.

Res. Lett., 33(12), L12603, doi:10.1029/2006GL026219, 2006.

Chen, C.-T. A., Huang, T.-H., Chen, Y.-C., Bai, Y., He, X. and Kang, Y.: Air–sea exchanges of CO_2 in the world's coastal seas, Biogeosciences, 10(10), 6509–6544, doi:10.5194/bg-10-6509-2013, 2013.

415   Christian, J. R., Arora, V. K., Boer, G. J., Curry, C. L., Zahariev, K., Denman, K. L., Flato, G. M., Lee, W. G., Merryfield, W. J., Roulet, N. T. and Scinocca, J. F.: The global carbon cycle in the Canadian Earth system model (CanESM1): Preindustrial control simulation, J. Geophys. Res., 115(G3), G03014, doi:10.1029/2008JG000920, 2010.

Claret, M., Galbraith, E. D., Palter, J. B., Bianchi, D., Fennel, K., Gilbert, D. and Dunne, J. P.: Rapid coastal deoxygenation due to ocean circulation shift in the northwest Atlantic, Nat. Clim. Chang., 8(10), 868–872, doi:10.1038/s41558-018-0263-1, 420   2018.

Collins, W. J., Bellouin, N., Doutriaux-Boucher, M., Gedney, N., Halloran, P., Hinton, T., Hughes, J., Jones, C. D., Joshi, M., Liddicoat, S., Martin, G., O'Connor, F., Rae, J., Senior, C., Sitch, S., Totterdell, I., Wiltshire, A. and Woodward, S.: Development and evaluation of an Earth-System model – HadGEM2, Geosci. Model Dev., 4(4), 1051–1075, doi:10.5194/gmd-4-1051-2011, 2011.

425   Danabasoglu, G., Lamarque, J. -F., Bacmeister, J., Bailey, D. A., DuVivier, A. K., Edwards, J., Emmons, L. K., Fasullo, J., Garcia, R., Gettelman, A., Hannay, C., Holland, M. M., Large, W. G., Lauritzen, P. H., Lawrence, D. M., Lenaerts, J. T. M., Lindsay, K., Lipscomb, W. H., Mills, M. J., Neale, R., Oleson, K. W., Otto-Bliesner, B., Phillips, A. S., Sacks, W., Tilmes, S., Kampenhout, L., Vertenstein, M., Bertini, A., Dennis, J., Deser, C., Fischer, C., Fox-Kemper, B., Kay, J. E., Kinnison, D., Kushner, P. J., Larson, V. E., Long, M. C., Mickelson, S., Moore, J. K., Nienhouse, E., Polvani, L., Rasch, P. J. and Strand, 430   W. G.: The Community Earth System Model Version 2 (CESM2), J. Adv. Model. Earth Syst., 12(2), doi:10.1029/2019MS001916, 2020.

Donlon, C. J., Martin, M., Stark, J., Roberts-Jones, J., Fiedler, E. and Wimmer, W.: The Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system, Remote Sens. Environ., 116, 140–158, doi:10.1016/j.rse.2010.10.017, 2012.

Ducklow, H. and McCallister, S.: The biogeochemistry of carbon dioxide in the coastal oceans, in The Sea, Vol 13: The global 435   coastal ocean. Multiscale interdisciplinary processes, edited by A. R. Robinson and K. H. Brink, pp. 269–315, John Wiley & Sons, New York., 2004.

Dufresne, J. L., Foujols, M. A., Denvil, S., Caubel, A., Marti, O., Aumont, O., Balkanski, Y., Bekki, S., Bellenger, H., Benshila, R., Bony, S., Bopp, L., Braconnot, P., Brockmann, P., Cadule, P., Cheruy, F., Codron, F., Cozic, A., Cugnet, D., de Noblet, N., Duvel, J. P., Ethé, C., Fairhead, L., Fichefet, T., Flavoni, S., Friedlingstein, P., Grandpeix, J. Y., Guez, L., Guilyardi, E., 440   Hauglustaine, D., Hourdin, F., Idelkadi, A., Ghattas, J., Joussaume, S., Kageyama, M., Krinner, G., Labetoulle, S., Lahellec, A., Lefebvre, M. P., Lefevre, F., Levy, C., Li, Z. X., Lloyd, J., Lott, F., Madec, G., Mancip, M., Marchand, M., Masson, S., Meurdesoif, Y., Mignot, J., Musat, I., Parouty, S., Polcher, J., Rio, C., Schulz, M., Swingedouw, D., Szopa, S., Talandier, C., Terray, P., Viovy, N. and Vuichard, N.: Climate change projections using the IPSL-CM5 Earth System Model: From CMIP3 to CMIP5., 2013.

445   Dunne, J. P.: Technical description of Tracers of Ocean Phytoplankton with Allometric Zooplankton version 2 (TOPAZ2)

used in GFDL's ESM2M and ESM2G submitted as part of the coupled model intercomparison project phase 5., 2013.

Dunne, J. P., John, J. G., Adcroft, A. J., Griffies, S. M., Hallberg, R. W., Shevliakova, E., Stouffer, R. J., Cooke, W., Dunne, K. A., Harrison, M. J., Krasting, J. P., Malyshev, S. L., Milly, P. C. D., Phillipps, P. J., Sentman, L. T., Samuels, B. L., Spelman, M. J., Winton, M., Wittenberg, A. T. and Zadeh, N.: GFDL's ESM2 Global Coupled Climate–Carbon Earth System Models.

450   Part I: Physical Formulation and Baseline Simulation Characteristics, J. Clim., 25(19), 6646–6665, doi:10.1175/JCLI-D-11-00560.1, 2012.

Dunne, J. P., John, J. G., Shevliakova, E., Stouffer, R. J., Krasting, J. P., Malyshev, S. L., Milly, P. C. D., Sentman, L. T., Adcroft, A. J., Cooke, W., Dunne, K. A., Griffies, S. M., Hallberg, R. W., Harrison, M. J., Levy, H., Wittenberg, A. T., Phillips, P. J. and Zadeh, N.: GFDL's ESM2 Global Coupled Climate–Carbon Earth System Models. Part II: Carbon System

455   Formulation and Baseline Simulation Characteristics*, J. Clim., 26(7), 2247–2267, doi:10.1175/JCLI-D-12-00150.1, 2013.

Dunne, J. P., Horowitz, L. W., Adcroft, A. J., Ginoux, P., Held, I. M., John, J. G., Krasting, J. P., Malyshev, S., Naik, V., Paulot, F., Shevliakova, E., A.Stock, C., Zadeh, N., Balaji, V., Blanton, C., Dunne, K. A., Dupuis, C., Durachta, J., Dussin, R., Gauthier, P. P. G., Griffies, S. M., Guo, H., Hallberg, R. W., Harrison, M., J. He, W. H., McHugh, C., Menzel, R., Milly, P. C. D., Nikonov, S., Paynter, D. J., Ploshay, J., Radhakrishnan, A., Rand, K., Reichl, B. G., Robinson, T., Schwarzkopf, D. M.,

460   Sentman, L. A., Underwood, S., Vahlenkamp, H., Winton, M., Wittenberg, A. T., Wyman, B., Zeng, Y. and Zhao, M.: The GFDL Earth System Model version 4.1 (GFDL-ESM4.1): Model description and simulation characteristics, J. Adv. Model. Earth Syst., 2019MS0020, n.d.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J. and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geosci. Model Dev., 9(5), 1937–

465   1958, doi:10.5194/gmd-9-1937-2016, 2016.

Fennel, K., Wilkin, J., Levin, J., Moisan, J., O'Reilly, J. and Haidvogel, D.: Nitrogen cycling in the Middle Atlantic Bight: Results from a three-dimensional model and implications for the North Atlantic nitrogen budget, Global Biogeochem. Cycles, 20(3), GB3007, doi:10.1029/2005GB002456, 2006.

Fennel, K., Wilkin, J., Previdi, M. and Najjar, R.: Denitrification effects on air-sea CO2 flux in the coastal ocean: Simulations

470   for the northwest North Atlantic, Geophys. Res. Lett., 35(24), L24608, doi:10.1029/2008GL036147, 2008.

Fennel, K., Hu, J., Laurent, A., Marta-Almeida, M. and Hetland, R.: Sensitivity of hypoxia predictions for the northern Gulf of Mexico to sediment oxygen consumption and model nesting, J. Geophys. Res. Ocean., 118(2), 990–1002, doi:10.1002/jgrc.20077, 2013.

Fennel, K., Alin, S., Barbero, L., Evans, W., Bourgeois, T., Cooley, S., Dunne, J., Feely, R. A., Hernandez-Ayon, J. M., Hu,

475   X., Lohrenz, S., Muller-Karger, F., Najjar, R., Robbins, L., Shadwick, E., Siedlecki, S., Steiner, N., Sutton, A., Turk, D., Vlahos, P. and Wang, Z. A.: Carbon cycling in the North American coastal ocean: a synthesis, Biogeosciences, 16(6), 1281–1304, doi:10.5194/bg-16-1281-2019, 2019.

Garcia, H. E., Locarnini, R. A., Boyer, T. P., Antonov, J. I., Zweng, M. M., Baranova, O. K. and Johnson, D. R.: World Ocean Atlas 2009, Volume 4: Nutrients (phosphate, nitrate, silicate), NOAA Atlas., edited by S. Levitus, U.S. Government Printing

480   Office, Washington, D.C., 2010.

Gilbert, D., Rabalais, N. N., Díaz, R. J. and Zhang, J.: Evidence for greater oxygen decline rates in the coastal ocean than in the open ocean, Biogeosciences, 7, 2283–2296, doi:10.5194/bg-7-2283-2010, 2010.

Giorgetta, M. A., Jungclaus, J., Reick, C. H., Legutke, S., Bader, J., Böttinger, M., Brovkin, V., Crueger, T., Esch, M., Fieg, K., Glushak, K., Gayler, V., Haak, H., Hollweg, H.-D., Ilyina, T., Kinne, S., Kornblueh, L., Matei, D., Mauritsen, T.,

485   Mikolajewicz, U., Mueller, W., Notz, D., Pithan, F., Raddatz, T., Rast, S., Redler, R., Roeckner, E., Schmidt, H., Schnur, R., Segschneider, J., Six, K. D., Stockhause, M., Timmreck, C., Wegner, J., Widmann, H., Wieners, K.-H., Claussen, M., Marotzke, J. and Stevens, B.: Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5, J. Adv. Model. Earth Syst., 5(3), 572–597, doi:10.1002/jame.20038, 2013.

Greenan, B. J. ., Petrie, B. ., Harrison, W. . and Oakey, N. .: Are the spring and fall blooms on the Scotian Shelf related to

490   short-term physical events?, Cont. Shelf Res., 24(4–5), 603–625, doi:10.1016/j.csr.2003.11.006, 2004.

Greenan, B. J. W., Petrie, B. D., Harrison, W. G. and Strain, P. M.: The onset and evolution of a spring bloom on the Scotian Shelf, , 53(5), 1759–1775, 2008.

Gruber, N., Hauri, C., Lachkar, Z., Loher, D., Frolicher, T. L. and Plattner, G.-K.: Rapid Progression of Ocean Acidification in the California Current System, Science (80-.)., 337(6091), 220–223, doi:10.1126/science.1216773, 2012.

495   Hajima, T., Watanabe, M., Yamamoto, A., Tatebe, H., Noguchi, M. A., Abe, M., Ohgaito, R., Ito, A., Yamazaki, D., Okajima, H., Ito, A., Takata, K., Ogochi, K., Watanabe, S. and Kawamiya, M.: Description of the MIROC-ES2L Earth system model and evaluation of its climate–biogeochemical processes and feedbacks, Rev., doi:10.5194/gmd-2019-275, 2019.

Hermann, A. J., Gibson, G. A., Bond, N. A., Curchitser, E. N., Hedstrom, K., Cheng, W., Wang, M., Cokelet, E. D., Stabeno, P. J. and Aydin, K.: Projected future biophysical states of the Bering Sea, Deep Sea Res. Part II Top. Stud. Oceanogr., 134,

500   30–47, doi:10.1016/j.dsr2.2015.11.001, 2016.

Hermann, A. J., Gibson, G. A., Cheng, W., Ortiz, I., Aydin, K., Wang, M., Hollowed, A. B. and Holsman, K. K.: Projected biophysical conditions of the Bering Sea to 2100 under multiple emission scenarios, ICES J. Mar. Sci., 76(6), 1937–1937, doi:10.1093/icesjms/fsz111, 2019.

Holt, J., Harle, J., Proctor, R., Michel, S., Ashworth, M., Batstone, C., Allen, I., Holmes, R., Smyth, T., Haines, K., Bretherton,

505   D. and Smith, G.: Modelling the global coastal ocean, Philos. Trans. R. Soc. A Math. Phys. Eng. Sci., 367(1890), 939–951, doi:10.1098/rsta.2008.0210, 2009.

Holt, J., Schrum, C., Cannaby, H., Daewel, U., Allen, I., Artioli, Y., Bopp, L., Butenschon, M., Fach, B. A., Harle, J., Pushpadas, D., Salihoglu, B. and Wakelin, S.: Potential impacts of climate change on the primary production of regional seas: A comparative analysis of five European seas, Prog. Oceanogr., 140, 91–115, doi:10.1016/j.pocean.2015.11.004, 2016.

510   Holt, J., Hyder, P., Ashworth, M., Harle, J., Hewitt, H. T., Liu, H., New, A. L., Pickles, S., Porter, A., Popova, E., Allen, J. I., Siddorn, J. and Wood, R.: Prospects for improving the representation of coastal and shelf seas in global ocean models, Geosci. Model Dev., 10(1), 499–523, doi:10.5194/gmd-10-499-2017, 2017.

Ilyina, T., Six, K. D., Segschneider, J., Maier-Reimer, E., Li, H. and Núñez-Riboni, I.: Global ocean biogeochemistry model

HAMOCC: Model architecture and performance as component of the MPI-Earth system model in different CMIP5

515 experimental realizations, J. Adv. Model. Earth Syst., 5(2), 287–315, doi:10.1029/2012MS000178, 2013.

Kuhn, A. M.: Integration of observations and models for an improved understanding of marine ecosystem dynamics, Dalhousie University. [online] Available from: http://hdl.handle.net/10222/73354, 2017.

Kuhn, A. M. and Fennel, K.: Evaluating ecosystem model complexity for the northwest North Atlantic through surrogate-based optimization, Ocean Model., 142, 101437, doi:10.1016/j.ocemod.2019.101437, 2019.

520 Kwiatkowski, L., Torres, O., Bopp, L., Aumont, O., Chamberlain, M., Christian, J. R., Dunne, J. P., Gehlen, M., Ilyina, T., John, J. G., Lenton, A., Li, H., Lovenduski, N. S., Orr, J. C., Palmieri, J., Santana-Falcón, Y., Schwinger, J., Séférian, R., Stock, C. A., Tagliabue, A., Takano, Y., Tjiputra, J., Toyama, K., Tsujino, H., Watanabe, M., Yamamoto, A., Yool, A. and Ziehn, T.: Twenty-first century ocean warming, acidification, deoxygenation, and upper-ocean nutrient and primary production decline from CMIP6 model projections, Biogeosciences, 17(13), 3439–3470, doi:10.5194/bg-17-3439-2020, 2020.

525 Lachkar, Z. and Gruber, N.: What controls biological production in coastal upwelling systems? Insights from a comparative modeling study, Biogeosciences, 8(10), 2961–2976, doi:10.5194/bg-8-2961-2011, 2011.

Laruelle, G. G., Cai, W. J., Hu, X., Gruber, N., Mackenzie, F. T. and Regnier, P.: Continental shelves as a variable but increasing global sink for atmospheric carbon dioxide, Nat. Commun., 9(1), doi:10.1038/s41467-017-02738-z, 2018.

Laurent, A., Fennel, K., Ko, D. S. and Lehrter, J.: Climate Change Projected to Exacerbate Impacts of Coastal Eutrophication

530 in the Northern Gulf of Mexico, J. Geophys. Res. Ocean., 123(5), 3408–3426, doi:10.1002/2017JC013583, 2018.

Lavoie, D., Lambert, N., den Mustapha, S. and van der Baaren, A.: Projections of Future Physical and Biogeochemical Conditions in the Northwest Atlantic from CMIP5 Global Climate Models., 2013.

Lavoie, D., Lambert, N. and Gilbert, D.: Projections of Future Trends in Biogeochemical Conditions in the Northwest Atlantic Using CMIP5 Earth System Models, Atmosphere-Ocean, 57(1), 18–40, doi:10.1080/07055900.2017.1401973, 2019.

535 Lindsay, K., Bonan, G. B., Doney, S. C., Hoffman, F. M., Lawrence, D. M., Long, M. C., Mahowald, N. M., Keith Moore, J., Randerson, J. T. and Thornton, P. E.: Preindustrial-Control and Twentieth-Century Carbon Cycle Experiments with the Earth System Model CESM1(BGC), J. Clim., 27(24), 8981–9005, doi:10.1175/JCLI-D-12-00565.1, 2014.

Loder, J. W., van der Baaren, A. and Yashayaev, I.: Climate Comparisons and Change Projections for the Northwest Atlantic from Six CMIP5 Models, Atmosphere-Ocean, 53(5), 529–555, doi:10.1080/07055900.2015.1087836, 2015.

540 Mattern, J. P. and Edwards, C. A.: Simple parameter estimation for complex models — Testing evolutionary techniques on 3-dimensional biogeochemical ocean models, J. Mar. Syst., 165, 139–152, doi:10.1016/j.jmarsys.2016.10.012, 2017.

Moore, J. K., Lindsay, K., Doney, S. C., Long, M. C. and Misumi, K.: Marine Ecosystem Dynamics and Biogeochemical Cycling in the Community Earth System Model [CESM1(BGC)]: Comparison of the 1990s with the 2090s under the RCP4.5 and RCP8.5 Scenarios, J. Clim., 26(23), 9291–9312, doi:10.1175/JCLI-D-12-00566.1, 2013.

545 Muller-Karger, F. E.: The importance of continental margins in the global carbon cycle, Geophys. Res. Lett., 32(1), L01602, doi:10.1029/2004GL021346, 2005.

Müller, W. A., Jungclaus, J. H., Mauritsen, T., Baehr, J., Bittner, M., Budich, R., Bunzel, F., Esch, M., Ghosh, R., Haak, H.,

Ilyina, T., Kleine, T., Kornblueh, L., Li, H., Modali, K., Notz, D., Pohlmann, H., Roeckner, E., Stemmler, I., Tian, F. and Marotzke, J.: A Higher-resolution Version of the Max Planck Institute Earth System Model (MPI-ESM1.2-HR), J. Adv. Model. 550 Earth Syst., 10(7), 1383–1413, doi:10.1029/2017MS001217, 2018.

Palmer, J. R. and Totterdell, I. J.: Production and export in a global ocean ecosystem model, Deep Sea Res. Part I Oceanogr. Res. Pap., 48(5), 1169–1198, doi:10.1016/S0967-0637(00)00080-7, 2001.

Peña, M. A., Fine, I. and Callendar, W.: Interannual variability in primary production and shelf-offshore transport of nutrients along the northeast Pacific Ocean margin, Deep Sea Res. Part II Top. Stud. Oceanogr., 169–170, 104637, 555 doi:10.1016/j.dsr2.2019.104637, 2019.

Rickard, G. J., Behrens, E. and Chiswell, S. M.: CMIP5 earth system models with biogeochemistry: An assessment for the southwest Pacific Ocean, J. Geophys. Res. Ocean., 121(10), 7857–7879, doi:10.1002/2016JC011736, 2016.

Romanou, A., Gregg, W. W., Romanski, J., Kelley, M., Bleck, R., Healy, R., Nazarenko, L., Russell, G., Schmidt, G. A., Sun, S. and Tausnev, N.: Natural air–sea flux of CO2 in simulations of the NASA-GISS climate model: Sensitivity to the physical 560 ocean model formulation, Ocean Model., 66, 26–44, doi:10.1016/j.ocemod.2013.01.008, 2013.

Ross, T., Craig, S. E., Comeau, A., Davis, R., Dever, M. and Beck, M.: Blooms and subsurface phytoplankton layers on the Scotian Shelf: Insights from profiling gliders, J. Mar. Syst., 172, 118–127, doi:10.1016/j.jmarsys.2017.03.007, 2017.

Rousseaux, C. S. and Gregg, W. W.: Recent decadal trends in global phytoplankton composition, Global Biogeochem. Cycles, 29(10), 1674–1688, doi:10.1002/2015GB005139, 2015.

565 Rutherford, K. and Fennel, K.: Diagnosing transit times on the northwestern North Atlantic continental shelf, Ocean Sci., 14(5), 1207–1221, doi:10.5194/os-14-1207-2018, 2018.

Saba, V. S., Griffies, S. M., Anderson, W. G., Winton, M., Alexander, M. A., Delworth, T. L., Hare, J. A., Harrison, M. J., Rosati, A., Vecchi, G. A. and Zhang, R.: Enhanced warming of the Northwest Atlantic Ocean under climate change, J. Geophys. Res. Ocean., 121(1), 118–132, doi:10.1002/2015JC011346, 2016.

570 Schmidt, G. A., Kelley, M., Nazarenko, L., Ruedy, R., Russell, G. L., Aleinov, I., Bauer, M., Bauer, S. E., Bhat, M. K., Bleck, R., Canuto, V., Chen, Y.-H., Cheng, Y., Clune, T. L., Del Genio, A., de Fainchtein, R., Faluvegi, G., Hansen, J. E., Healy, R. J., Kiang, N. Y., Koch, D., Lacis, A. A., LeGrande, A. N., Lerner, J., Lo, K. K., Matthews, E. E., Menon, S., Miller, R. L., Oinas, V., Oloso, A. O., Perlwitz, J. P., Puma, M. J., Putman, W. M., Rind, D., Romanou, A., Sato, M., Shindell, D. T., Sun, S., Syed, R. A., Tausnev, N., Tsigaridis, K., Unger, N., Voulgarakis, A., Yao, M.-S. and Zhang, J.: Configuration and 575 assessment of the GISS ModelE2 contributions to the CMIP5 archive, J. Adv. Model. Earth Syst., 6(1), 141–184, doi:10.1002/2013MS000265, 2014.

Schneider, B., Bopp, L., Gehlen, M., Segschneider, J., Frölicher, T. L., Cadule, P., Friedlingstein, P., Doney, S. C., Behrenfeld, M. J. and Joos, F.: Climate-induced interannual variability of marine primary and export production in three global coupled climate carbon cycle models, Biogeosciences, 5(2), 597–614, doi:10.5194/bg-5-597-2008, 2008.

580 Séférian, R., Nabat, P., Michou, M., Saint-Martin, D., Voldoire, A., Colin, J., Decharme, B., Delire, C., Berthet, S., Chevallier, M., Sénési, S., Franchisteguy, L., Vial, J., Mallet, M., Joetzjer, E., Geoffroy, O., Guérémy, J., Moine, M., Msadek, R., Ribes,

A., Rocher, M., Roehrig, R., Salas-y-Mélia, D., Sanchez, E., Terray, L., Valcke, S., Waldman, R., Aumont, O., Bopp, L., Deshayes, J., Éthé, C. and Madec, G.: Evaluation of CNRM Earth System Model, CNRM-ESM2-1: Role of Earth System Processes in Present-Day and Future Climate, J. Adv. Model. Earth Syst., 2019MS001791, doi:10.1029/2019MS001791, 2019.

585 Seitzinger, S. P. and Giblin, A. E.: Estimating denitrification in North Atlantic continental shelf sediments, Biogeochemistry, 35(1), 235–260, doi:10.1007/BF02179829, 1996.

Sellar, A. A., Jones, C. G., Mulcahy, J., Tang, Y., Yool, A., Wiltshire, A., O'Connor, F. M., Stringer, M., Hill, R., Palmieri, J., Woodward, S., Mora, L., Kuhlbrodt, T., Rumbold, S., Kelley, D. I., Ellis, R., Johnson, C. E., Walton, J., Abraham, N. L., Andrews, M. B., Andrews, T., Archibald, A. T., Berthou, S., Burke, E., Blockley, E., Carslaw, K., Dalvi, M., Edwards, J.,

590 Folberth, G. A., Gedney, N., Griffiths, P. T., Harper, A. B., Hendry, M. A., Hewitt, A. J., Johnson, B., Jones, A., Jones, C. D., Keeble, J., Liddicoat, S., Morgenstern, O., Parker, R. J., Predoi, V., Robertson, E., Siahaan, A., Smith, R. S., Swaminathan, R., Woodhouse, M. T., Zeng, G. and Zerroukat, M.: UKESM1: Description and evaluation of the UK Earth System Model, J. Adv. Model. Earth Syst., 2019MS001739, doi:10.1029/2019MS001739, 2019.

Siedlecki, S. A., Banas, N. S., Davis, K. A., Giddings, S., Hickey, B. M., MacCready, P., Connolly, T. and Geier, S.: Seasonal

595 and interannual oxygen variability on the Washington and Oregon continental shelves, J. Geophys. Res. Ocean., 120(2), 608– 633, doi:10.1002/2014JC010254, 2015.

Swart, N. C., Cole, J. N. S., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., Anstey, J., Arora, V., Christian, J. R., Hanna, S., Jiao, Y., Lee, W. G., Majaess, F., Saenko, O. A., Seiler, C., Seinen, C., Shao, A., Solheim, L., von Salzen, K., Yang, D. and Winter, B.: The Canadian Earth System Model version 5 (CanESM5.0.3), Geosci. Model Dev. Discuss., 1–68,

600 doi:10.5194/gmd-2019-177, 2019.

Taylor, K. E., Stouffer, R. J. and Meehl, G. A.: An Overview of CMIP5 and the Experiment Design, Bull. Am. Meteorol. Soc., 93(4), 485–498, doi:10.1175/BAMS-D-11-00094.1, 2012.

Tjiputra, J. F., Roelandt, C., Bentsen, M., Lawrence, D. M., Lorentzen, T., Schwinger, J., Seland, Ø. and Heinze, C.: Evaluation of the carbon cycle components in the Norwegian Earth System Model (NorESM), Geosci. Model Dev., 6(2), 301–325,

605 doi:10.5194/gmd-6-301-2013, 2013.

Urrego-Blanco, J. and Sheng, J.: Interannual Variability of the Circulation over the Eastern Canadian Shelf, Atmosphere-Ocean, 50(3), 277–300, doi:10.1080/07055900.2012.680430, 2012.

Vichi, M., Pinardi, N. and Masina, S.: A generalized model of pelagic biogeochemistry for the global ocean ecosystem. Part I: Theory, J. Mar. Syst., 64(1–4), 89–109, doi:10.1016/j.jmarsys.2006.03.006, 2007a.

610 Vichi, M., Masina, S. and Navarra, a.: A generalized model of pelagic biogeochemistry for the global ocean ecosystem. Part II: Numerical simulations, J. Mar. Syst., 64(1–4), 110–134, doi:10.1016/j.jmarsys.2006.03.014, 2007b.

Vichi, M., Manzini, E., Fogli, P. G., Alessandri, A., Patara, L., Scoccimarro, E., Masina, S. and Navarra, A.: Global and regional ocean carbon uptake and climate change: sensitivity to a substantial mitigation scenario, Clim. Dyn., 37(9–10), 1929– 1947, doi:10.1007/s00382-011-1079-0, 2011.

615 Voldoire, A., Sanchez-Gomez, E., Salas y Mélia, D., Decharme, B., Cassou, C., Sénési, S., Valcke, S., Beau, I., Alias, A.,

Chevallier, M., Déqué, M., Deshayes, J., Douville, H., Fernandez, E., Madec, G., Maisonnave, E., Moine, M.-P., Planton, S., Saint-Martin, D., Szopa, S., Tyteca, S., Alkama, R., Belamari, S., Braun, A., Coquart, L. and Chauvin, F.: The CNRM-CM5.1 global climate model: description and basic evaluation, Clim. Dyn., 40(9–10), 2091–2121, doi:10.1007/s00382-011-1259-y, 2013.

620   Walsh, J. J.: Importance of continental margins in the marine biogeochemical cycling of carbon and nitrogen, Nature, 350, 53–55, 1991.

Yool, A., Popova, E. E. and Anderson, T. R.: MEDUSA-2.0: an intermediate complexity biogeochemical model of the marine carbon cycle for climate change and ocean acidification studies, Geosci. Model Dev., 6(5), 1767–1811, doi:10.5194/gmd-6-1767-2013, 2013.

625   Zhang, H., Fennel, K., Laurent, A. and Bian, C.: A numerical model study of the main factors contributing to hypoxia and its sub-seasonal to interannual variability off the Changjiang Estuary, Biogeosciences Discuss., Rev., doi:10.5194/bg-2019-341, 2019.

630   **Table 1. Information about the regional model and the 29 ESM models. For the CMIP5 models (2–18) the r1i1p1 ensemble was used. For the CMIP6 model (19–30) the r1i1p1f1 ensemble was used on the native grid when available, except for CNRM-ESM2-1, MIROC-ES2L and UKESM1-0-LL (r1i1p1f2), GFDL-CM4, GFDL-ESM4 and NorESM2-LM (regridded), and GISS-E2-1-G (r101i1p1f1).**

Model		Shelf resolution (n cells)			Ocean BGC component	References
Name	ID	GM	SS	GB		
ACM	1	1780	1366	3729	BIO_FENNEL	Brennan et al. (2016); Fennel et al. (2006)
CanESM2	2	11	14	29	CMOC	Arora et al. (2011); Christian et al. (2010)
CESM1-BGC	3	41	33	91	BEC	Lindsay et al. (2014); Moore et al. (2013)
CMCC-CESM	4	8	5	13	PELAGOS	Vichi et al. (2007a, 2007b, 2011)
CNRM-CM5	5	27	20	55	PISCES	Aumont and Bopp (2006); Voldoire et al. (2013)
GFDL-ESM2-G	6	20	15	39	TOPAZ2	Dunne (2013); Dunne et al. (2012, 2013)
GFDL-ESM2-M	7					
GISS-E2-H-CC	8	19	14	39	NOBM	Romanou et al. (2013) Schmidt et al. (2014)
GISS-E2-R-CC	9	15	12	29		
HadGEM2-CC	10	18	15	39	Diat-HadOCC	Collins et al., (2011); Palmer and Totterdell (2001)
HadGEM2-ES	11					
IPSL-CM5A-LR	12	8	5	13	PISCES	Aumont and Bopp (2006); Dufresne et al. (2013)
IPSL-CM5A-MR	13					
IPSL-CM5B-LR	14					
MPI-ESM-LR	15	23	23	73	HAMOCC 5.2	Giorgetta et al. (2013); Ilyina et al. 2013)
MPI-ESM-MR	16	136	87	193		
MRI-ESM1	17	40	29	80	MRI.COM3	Adachi et al. (2013)
NorESM1-ME	18	41	33	91	HAMOCC 5.1	Tjiputra et al. (2013)
CanESM5	19	27	20	55	CMOC	Swart et al. (2019)
CESM2	20	41	33	91	MARBL	Danabasoglu et al. (2020)
CESM2-WACCM	21					
CNRM-ESM2-1	22	27	20	55	PISCES	Aumont et al. (2015); Séférian et al. (2019)
GFDL-ESM4	23	20	15	39	COBALTv2	Dunne et al. (submitted)
GISS-E2-1-G	24	15	12	29	NOBM	Rousseaux and Gregg (2015)
GISS-E2-1-G-CC	25					
IPSL-CM6A-LR	26	27	20	55	PISCES	Aumont et al. (2015); Boucher et al. (2020)
MIROC-ES2L	27	20	18	43	OECO2	Hajima et al. (2019)
MPI-ESM1-2-HR	28	136	87	193	HAMOCC	Müller et al. (2018)
NorESM2-LM	29	25	20	57	HAMOCC	Müller et al. (2018)

UKESM1-0-LL	30	27	20	55	MEDUSA2	Sellar et al. (2019); Yool et al. (2013)

635 **Table 2. Annual model scores and ranking.**

Ranked models			Scores			Ranks					
Name	ID	CMIP	Temp.	Chl-a	NO3	Temp.	Chl-a	NO3	$\bar{R}$	$\bar{R}_{bio}$	Overall
ACM	1	–	1.14	0.64	1.27	1	1	1	1.0	1.0	1
MPI-ESM1-2-HR	28	6	2.05	1.03	1.75	4	7	6	5.7	6.5	2
CNRM-ESM2-1	22	6	2.74	0.92	1.91	12	3	9	8.0	6.0	3
GFDL-ESM4	23	6	2.49	0.81	2.10	9	2	13	8.0	7.5	4
IPSL-CM6A-LR	26	6	2.47	1.09	1.94	8	12	10	10.0	11.0	5
GFDL-ESM2G	6	5	2.12	1.17	1.67	5	20	5	10.0	12.5	6
HadGEM2-CC	10	5	2.58	1.02	2.11	11	6	14	10.3	10.0	7
IPSL-CM5A-MR	13	5	3.07	1.09	1.80	16	13	7	12.0	10.0	8
IPSL-CM5A-LR	12	5	2.52	1.17	1.91	10	19	8	12.3	12.0	9
GISS-E2-1-G-CC	25	6	4.66	1.08	1.44	25	11	2	12.7	13.5	10
HadGEM2-ES	11	5	2.90	1.06	2.12	14	9	15	12.7	6.5	11
GISS-E2-1-G	24	6	4.79	1.08	1.47	26	10	3	13.0	6.5	12
MIROC-ES2L	27	6	3.14	1.02	2.17	18	5	17	13.3	11.0	13
IPSL-CM5B-LR	14	5	1.51	1.36	2.03	2	26	12	13.3	19.0	14
MPI-ESM-MR	16	5	2.14	1.09	2.57	6	14	21	13.7	17.5	15
UKESM1-0-LL	30	6	3.08	1.15	1.96	17	17	11	15.0	14.0	16
MPI-ESM-LR	15	5	2.38	1.10	3.12	7	15	24	15.3	19.5	17
CNRM-CM5	5	5	1.78	1.11	6.54	3	16	27	15.3	21.5	18
GISS-E2-R-CC	9	5	3.84	1.19	1.62	21	22	4	15.7	13.0	19
MRI-ESM1	17	5	2.78	1.15	2.53	13	18	20	17.0	19.0	20
NorESM2-LM	29	6	2.98	1.04	13.26	15	8	30	17.7	19.0	21
GFDL-ESM2M	7	5	3.89	0.95	7.14	22	4	29	18.3	16.5	22
GISS-E2-H-CC	8	5	3.64	1.35	2.29	19	25	18	20.7	21.5	23
CanESM5	19	6	4.05	1.35	2.16	23	24	16	21.0	20.0	24
CanESM2	2	5	4.20	1.18	3.14	24	21	25	23.3	23.0	25
CMCC-CESM	4	5	5.18	1.40	2.39	27	29	19	25.0	24.0	26
NorESM1-ME	18	5	3.71	1.40	6.99	20	28	28	25.3	28.0	27
CESM2	20	6	5.40	1.38	2.61	28	27	22	25.7	24.5	28
CESM1-BGC	3	5	7.84	1.29	4.21	30	23	26	26.3	24.5	29
CESM2-WACCM	21	6	5.71	1.42	2.78	29	30	23	27.3	26.5	30

Table 3. Comparison statistics between ACM and AZMP and glider observations along the Halifax Line and at Station 2.

Season*	RMSD				Bias				Correlation coefficient			
	W	S	S	F	W	S	S	F	W	S	S	F
						Halifax Line						
Chlorophyll (time series)	0.25	0.37	0.39	0.36	0.08	0.22	0.28	0.13	0.68	0.78	0.71	0.75
Chlorophyll (Glider)	0.22	0.42	0.25	0.22	-0.14	0.13	0.17	0.04	0.88	0.78	0.94	0.85
Nitrate	2.99	2.73	2.13	1.77	0.76	2.03	0.74	1.27	0.90	0.83	0.85	0.92
						Station 2						
Chlorophyll (time series)	0.26	1.74	0.52	0.30	0.05	-0.56	0.26	0.01	0.64	0.22	0.48	0.82
Chlorophyll (Glider)	0.15	1.06	0.31	0.17	-0.03	-0.46	0.25	0.02	0.87	0.69	0.91	0.93
Nitrate	0.96	1.57	1.58	1.37	1.19	1.62	0.26	0.58	0.85	0.86	0.91	0.94

*Seasons are order sequentially and abbreviated as W (winter, Dec–Feb), S (spring, Mar–May), S (summer, Jun–Aug) and F (fall, Sep–Nov).

640
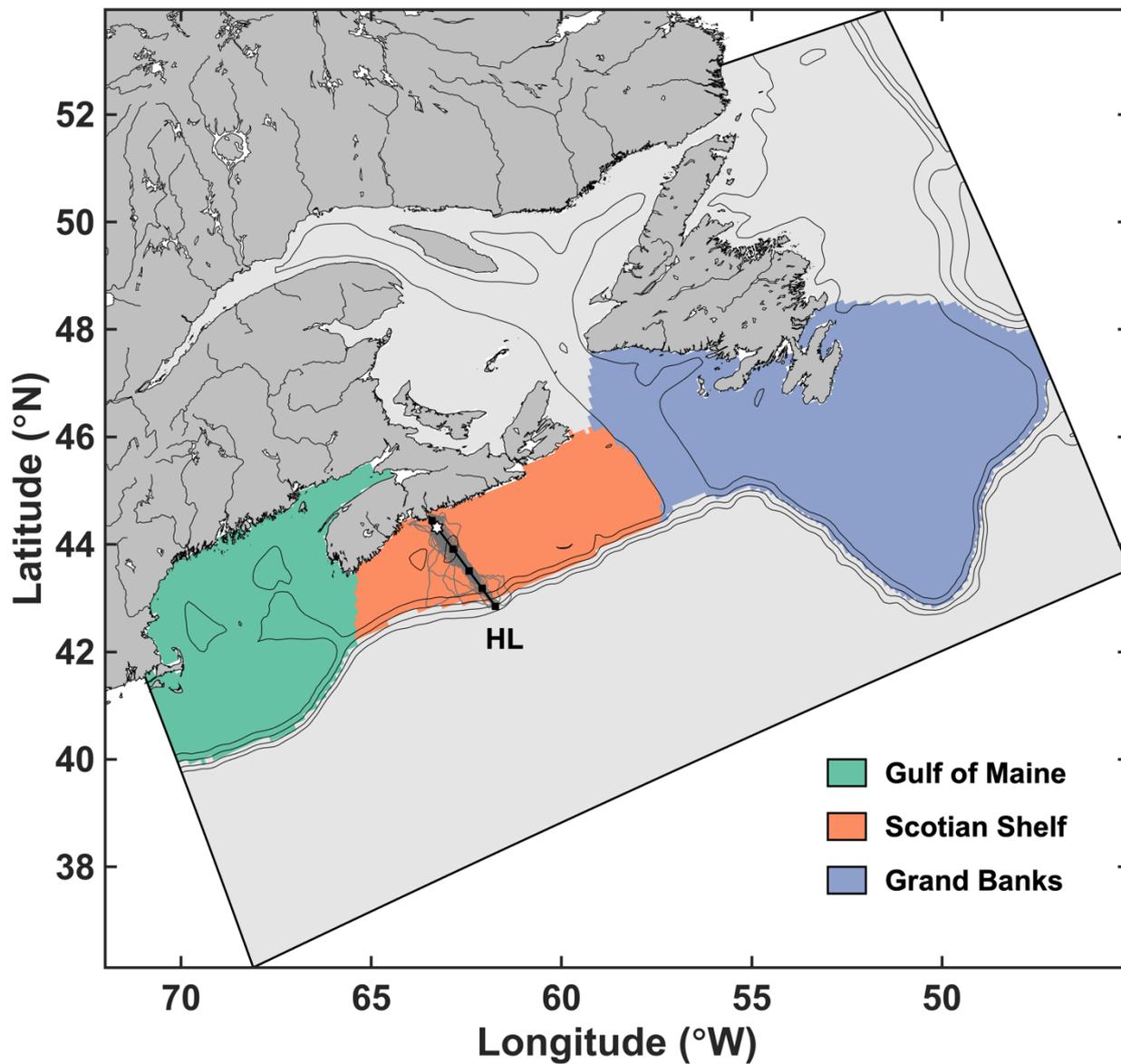
Figure 1. Study area indicating the 3 averaging zones, the limits of the ROMS grid and the location of the Halifax Line stations (squares) used in the analysis. The white star is Station 2 and the grey lines the gliders track.
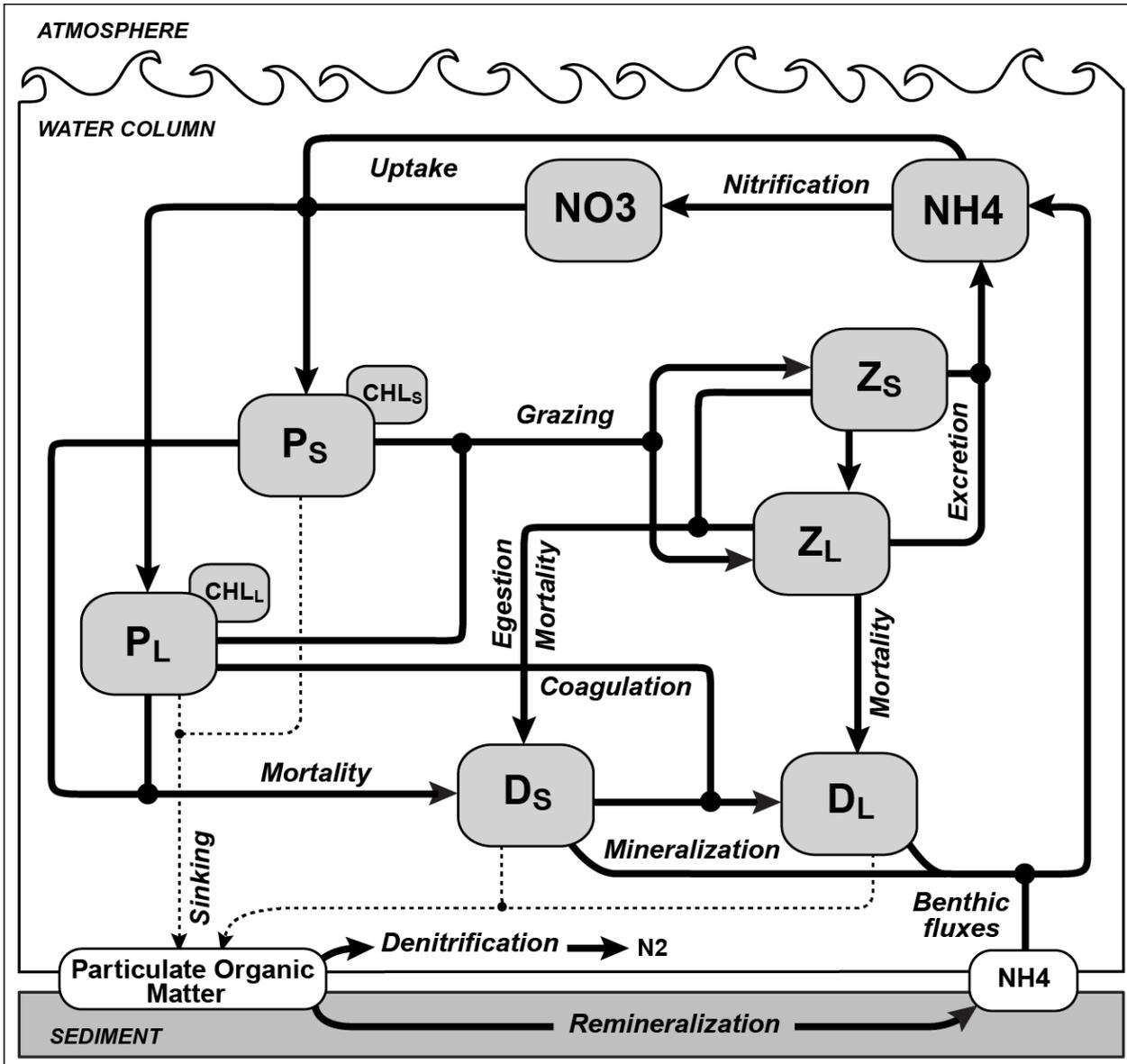
Figure 2. Schematic of the biogeochemical model used in ROMS. The state variables are small phytoplankton (P_S) and chlorophyll (CHL_S), large phytoplankton (P_L) and chlorophyll (CHL_L), small zooplankton (Z_S), large zooplankton (Z_L), slow-sinking small detritus (D_S), fast-sinking large detritus (D_L), nitrate (NO_3), and ammonium (NH_4). Dashed lines indicate sinking. Black dots represent the connections between paths.
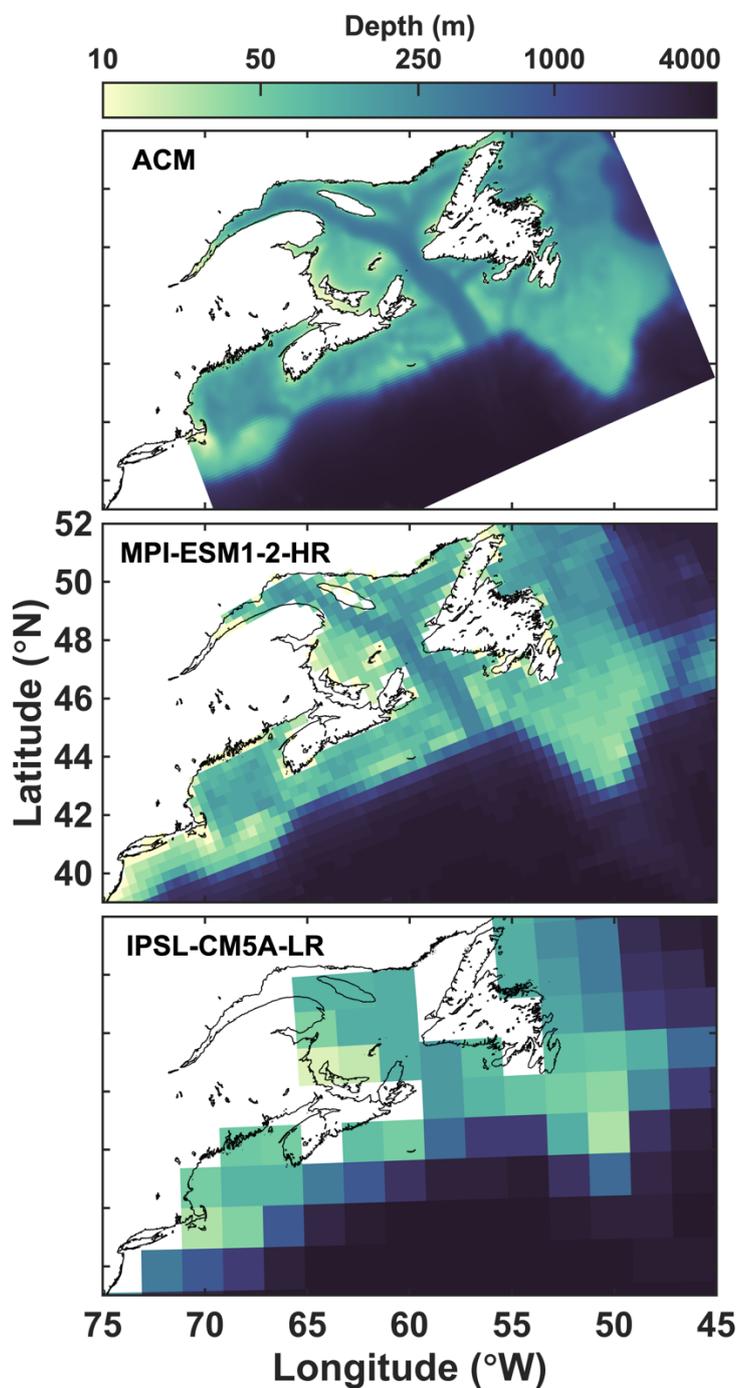
Figure 3. Bathymetry of the regional model (top), the highest resolution ESM (middle) and lowest resolution ESM (bottom).
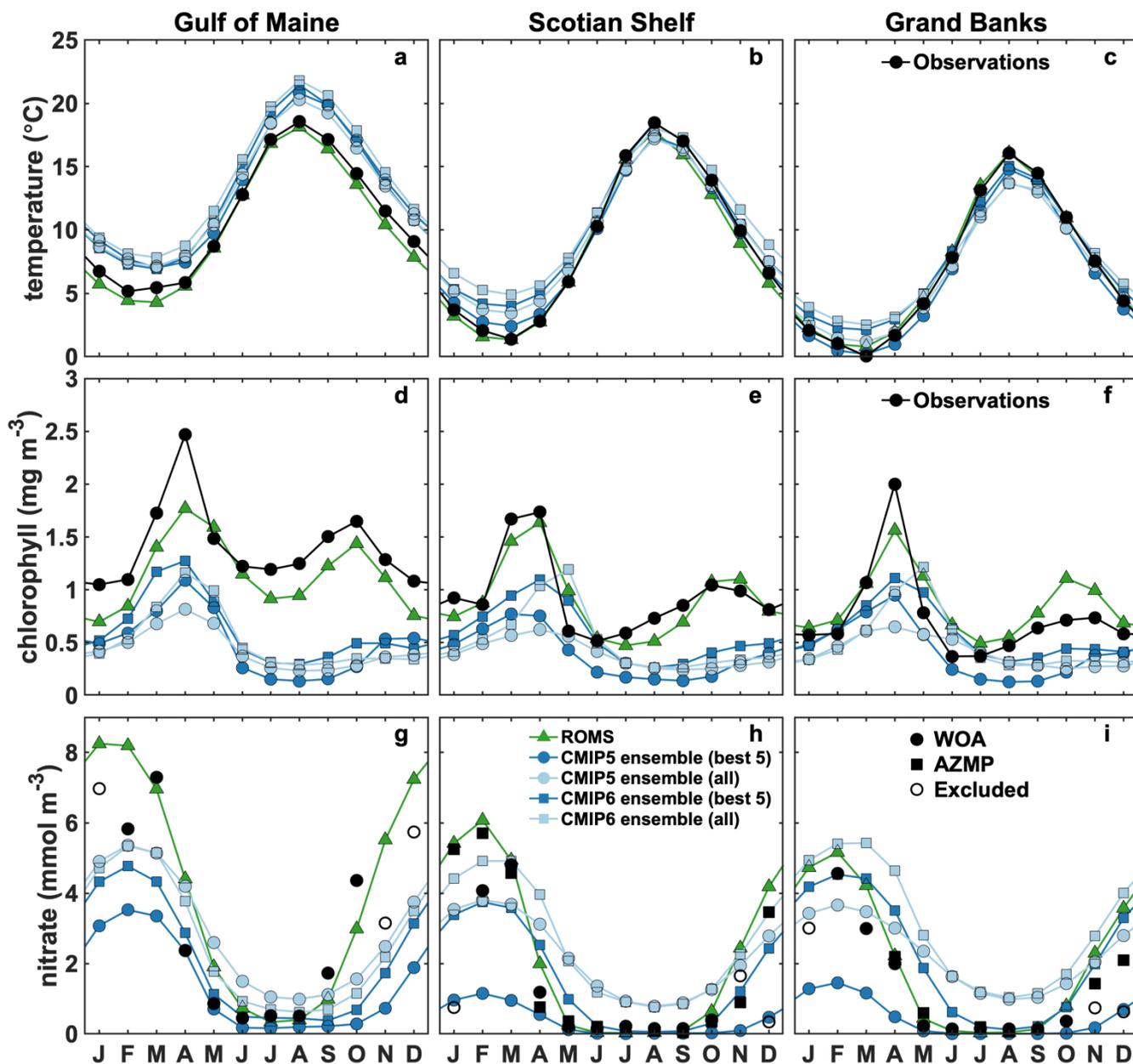
Figure 4. Observed, ROMS and ensemble means area averaged surface chlorophyll (a-c), nitrate (d-f) and temperature (g-i) in the 3 NWA shelf regions.
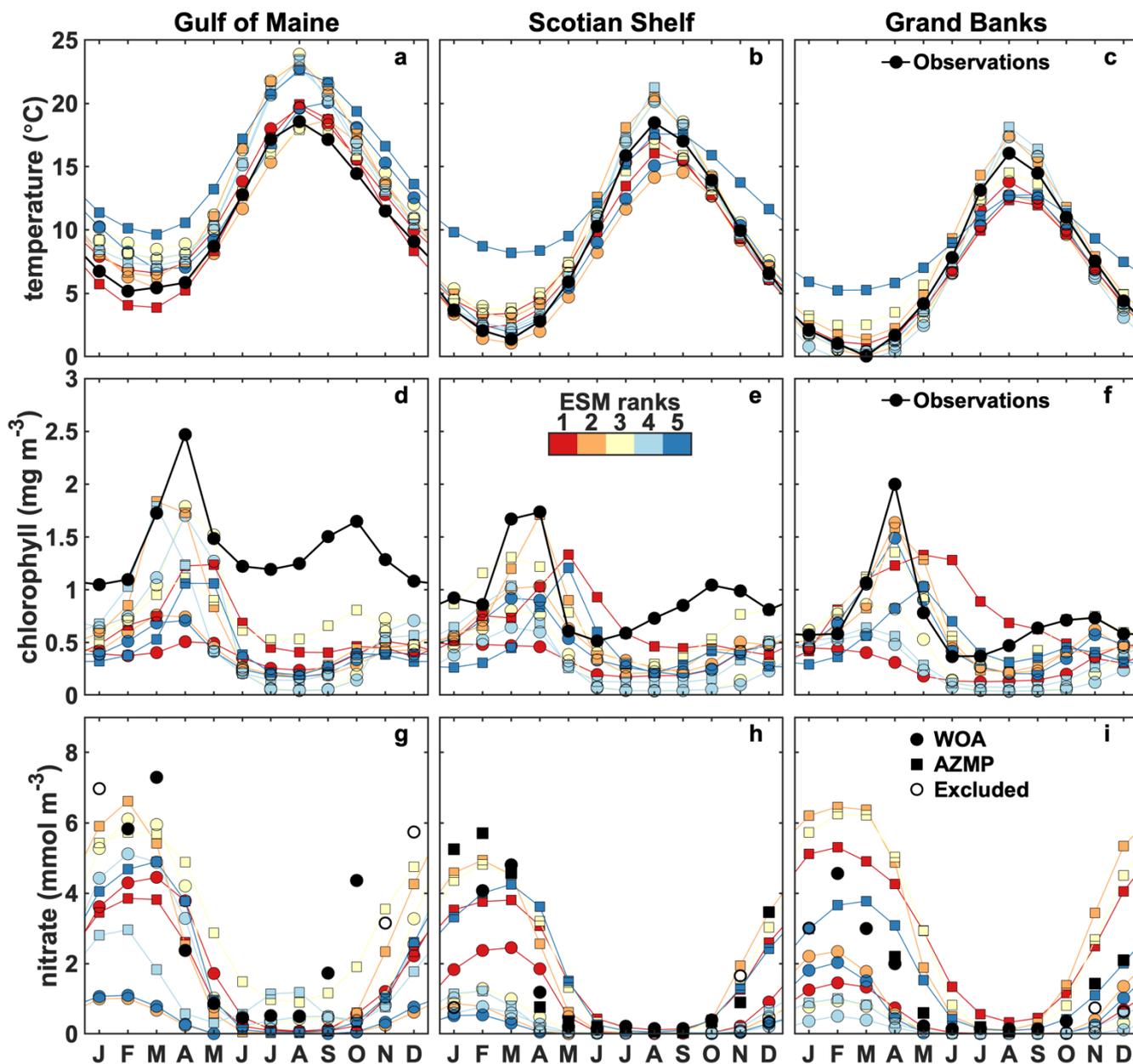
Figure 5. Observed (black dots) and best ESMs area averaged surface chlorophyll (a-c), nitrate (d-f) and temperature (g-i) in the 3 NWA shelf regions. The colored circles and squares indicate the CMIP5 and CMIP6 models, respectively.
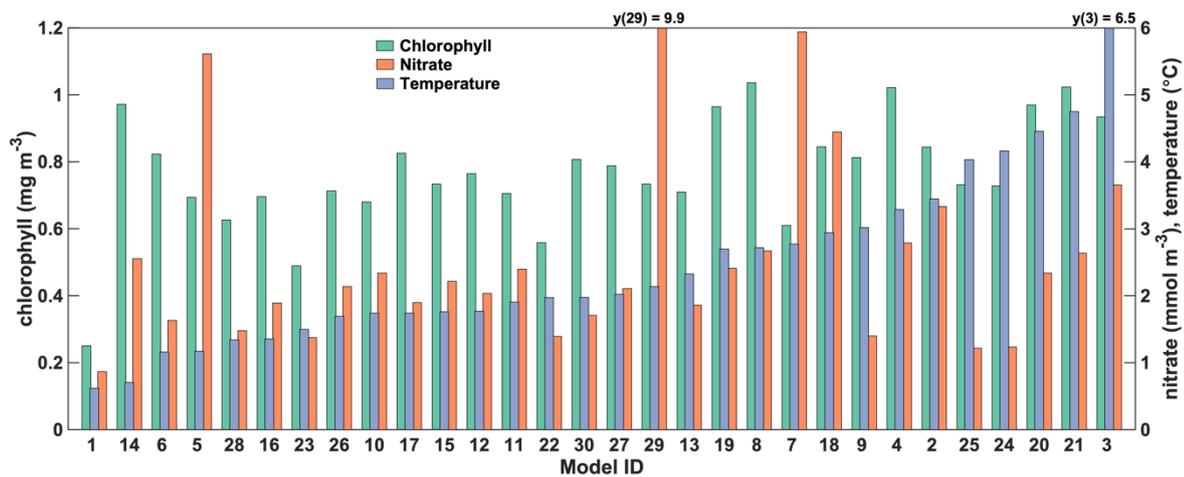
Figure 6. Root mean square difference between monthly, regionally averaged observations and models. Model numbers refer to the IDs in Table 1.

665
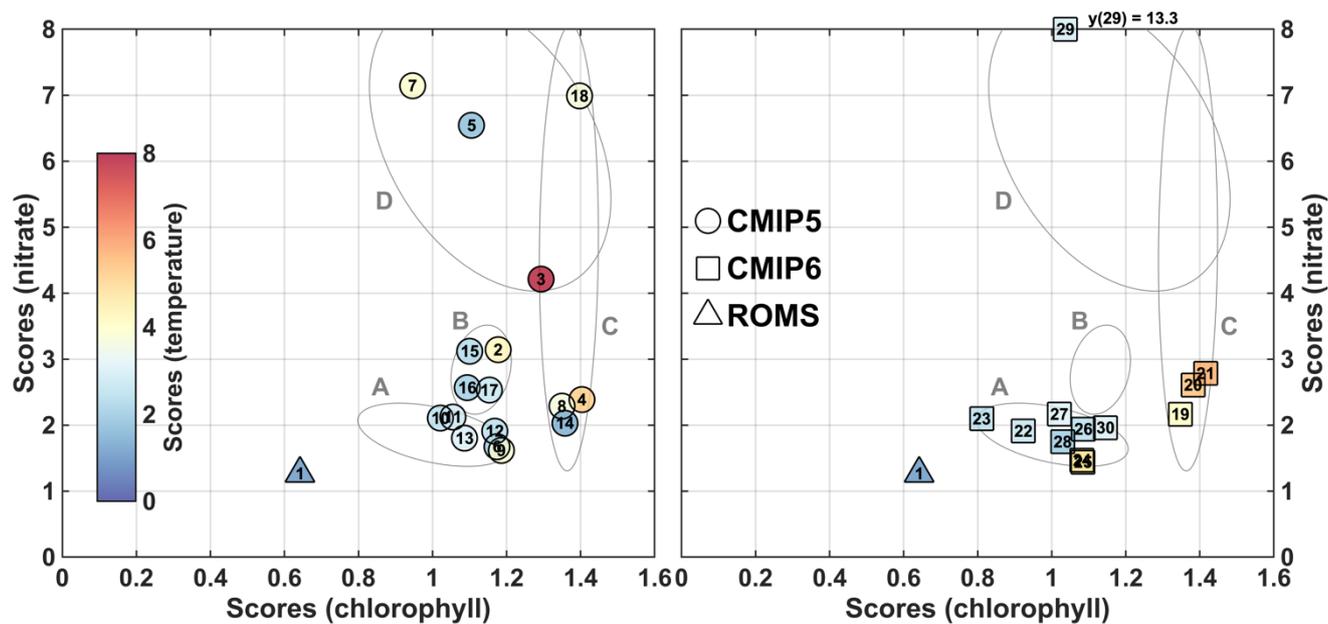
Figure 7. Model scores for surface chlorophyll (x-axis), nitrate (y-axis) and temperature (color scale) for the CMIP5 group (A, left panel), the CMIP6 group (B, right panel) and the regional model. The grey ellipsoids indicate the groups A–D (see text) and are the same in panels A and B.

670
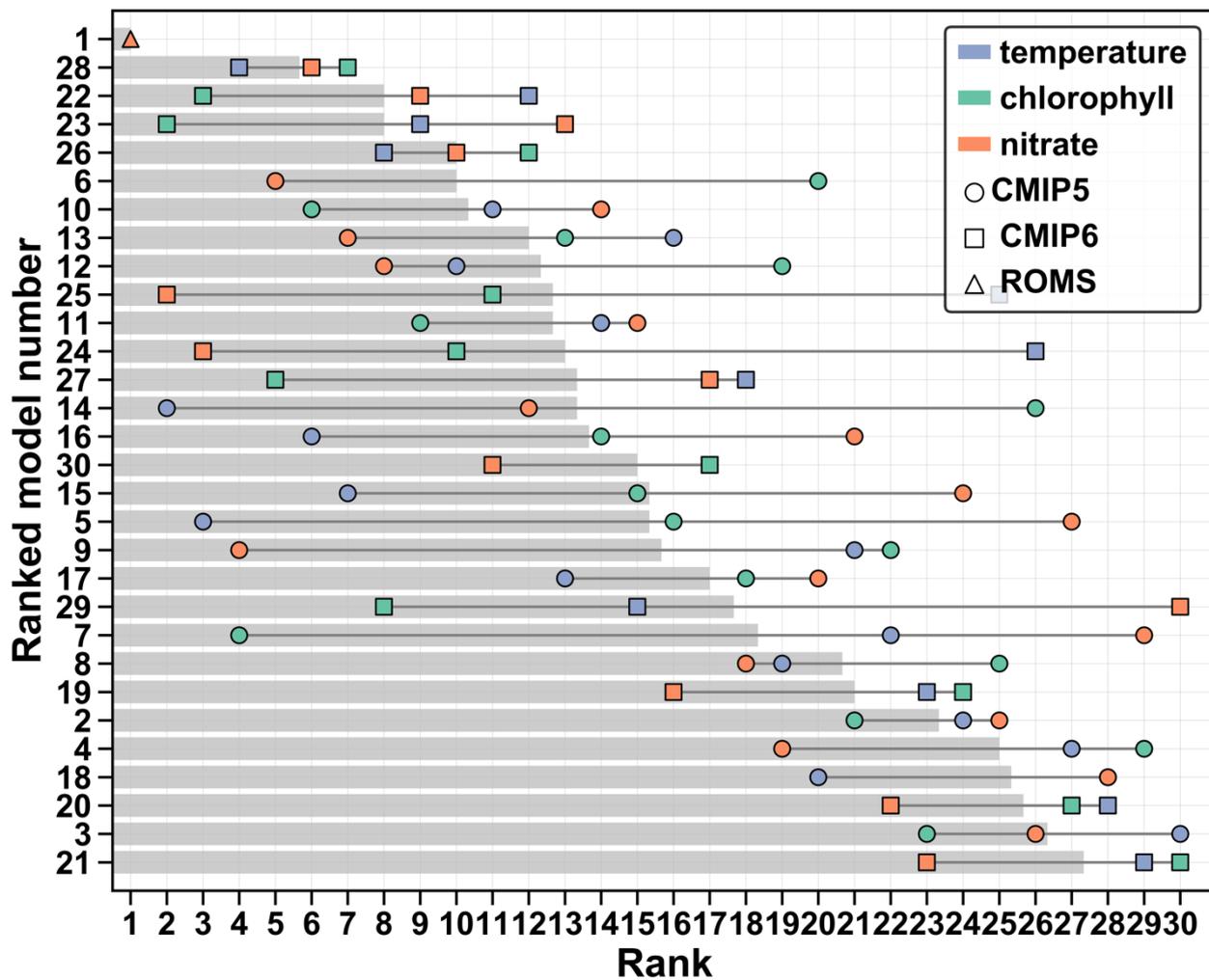
Figure 8. Model average (grey bars) and specific (dots) ranking. The final ranking is shown on the y-axis. The temperature rank for model 6 is hidden behind the nitrate rank (x = 5).

675
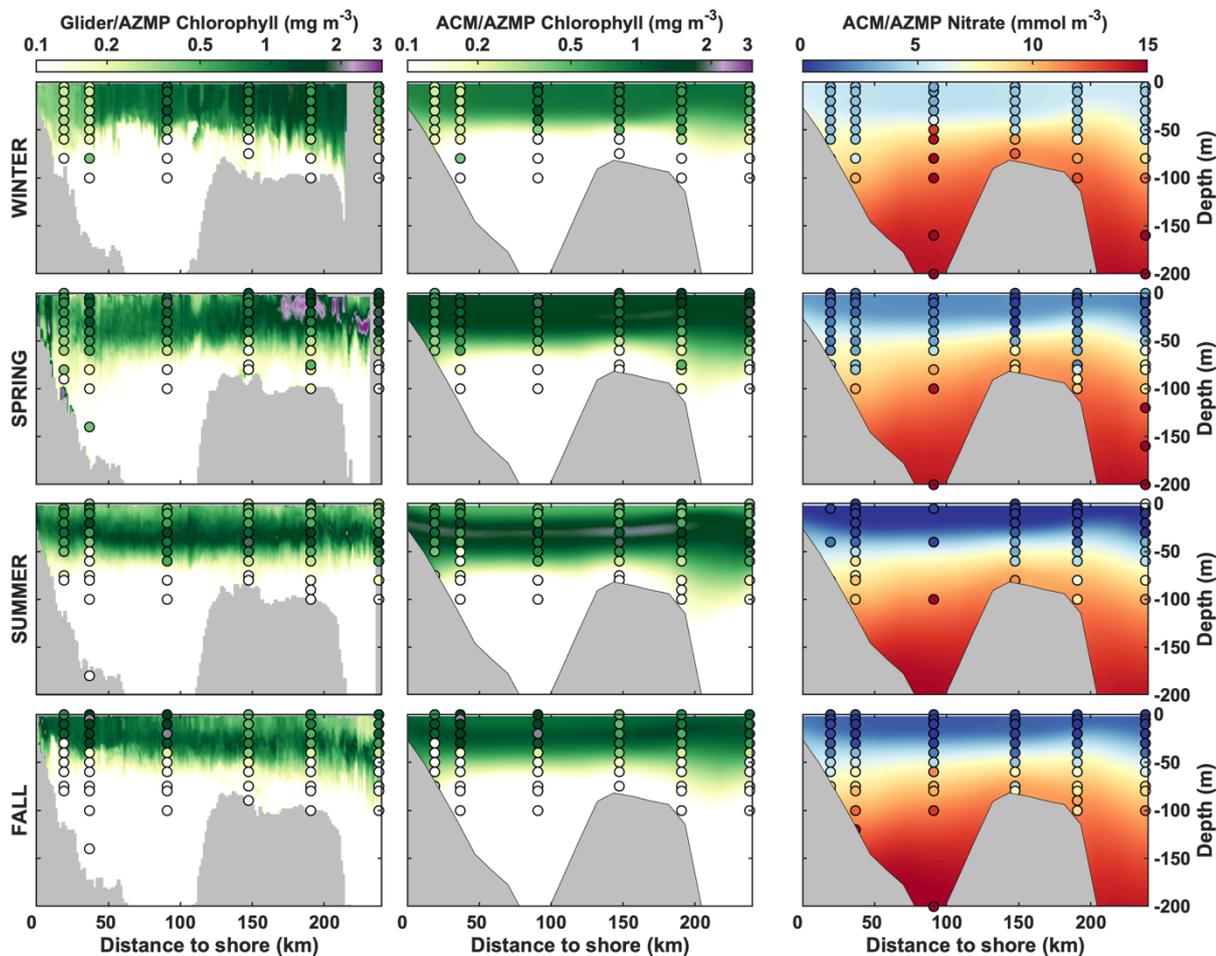
Figure 9. Comparison of gliders, AZMP and model seasonal climatologies of chlorophyll and nitrate along the Halifax line.
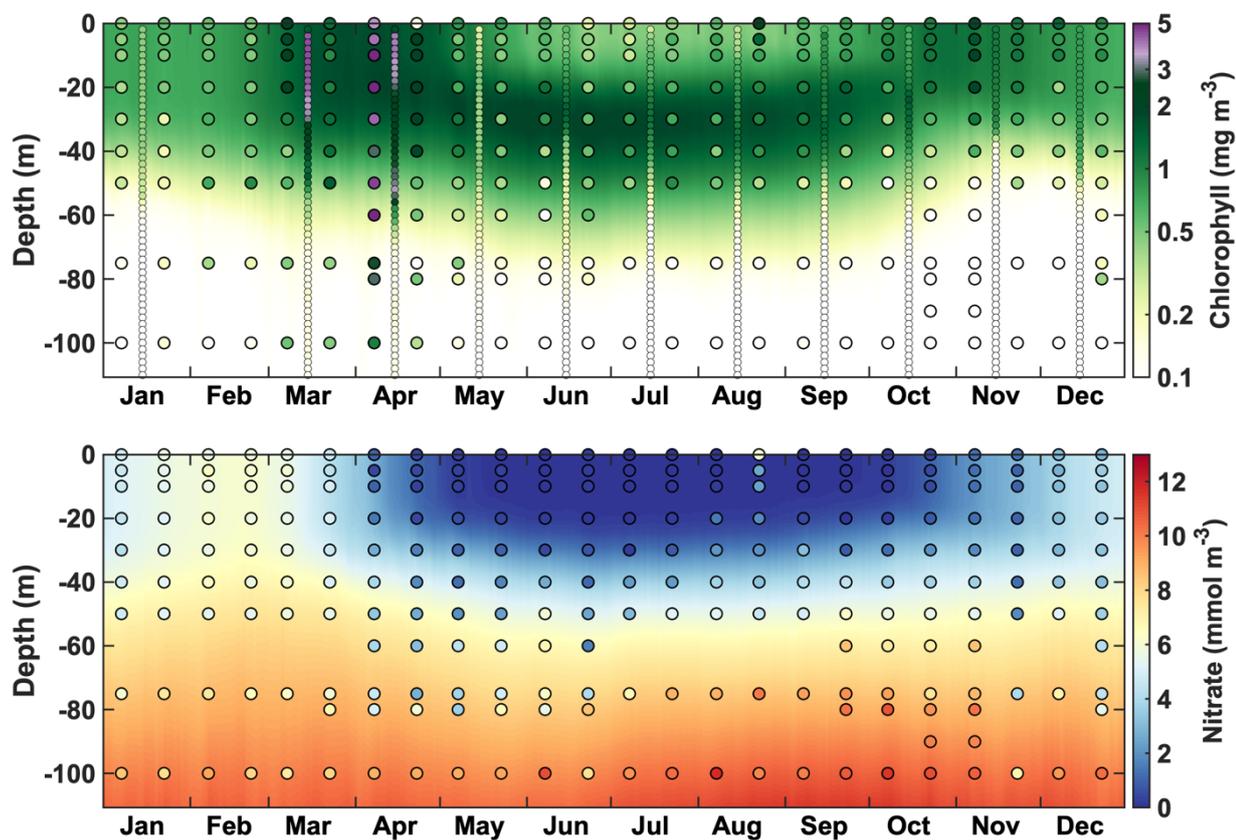
Figure 10. Comparison of vertically-resolved time series of chlorophyll (top) and nitrate (bottom) at Station 2 from the regional model (background), the glider transects (small dots) and the bimonthly sampling (large dots).

680
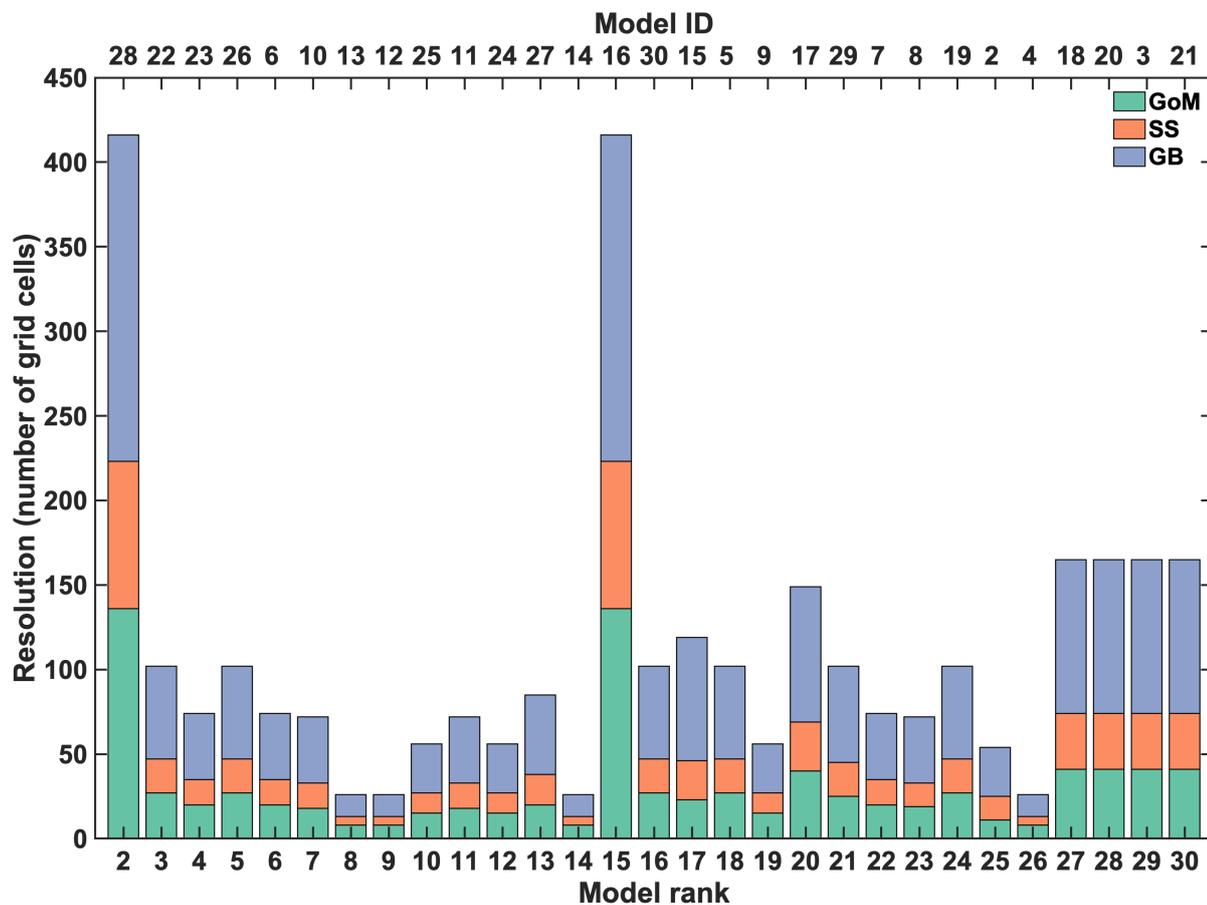
Figure 11. Resolution of the 29 ESMs ordered by their overall rank (see Figure 8).
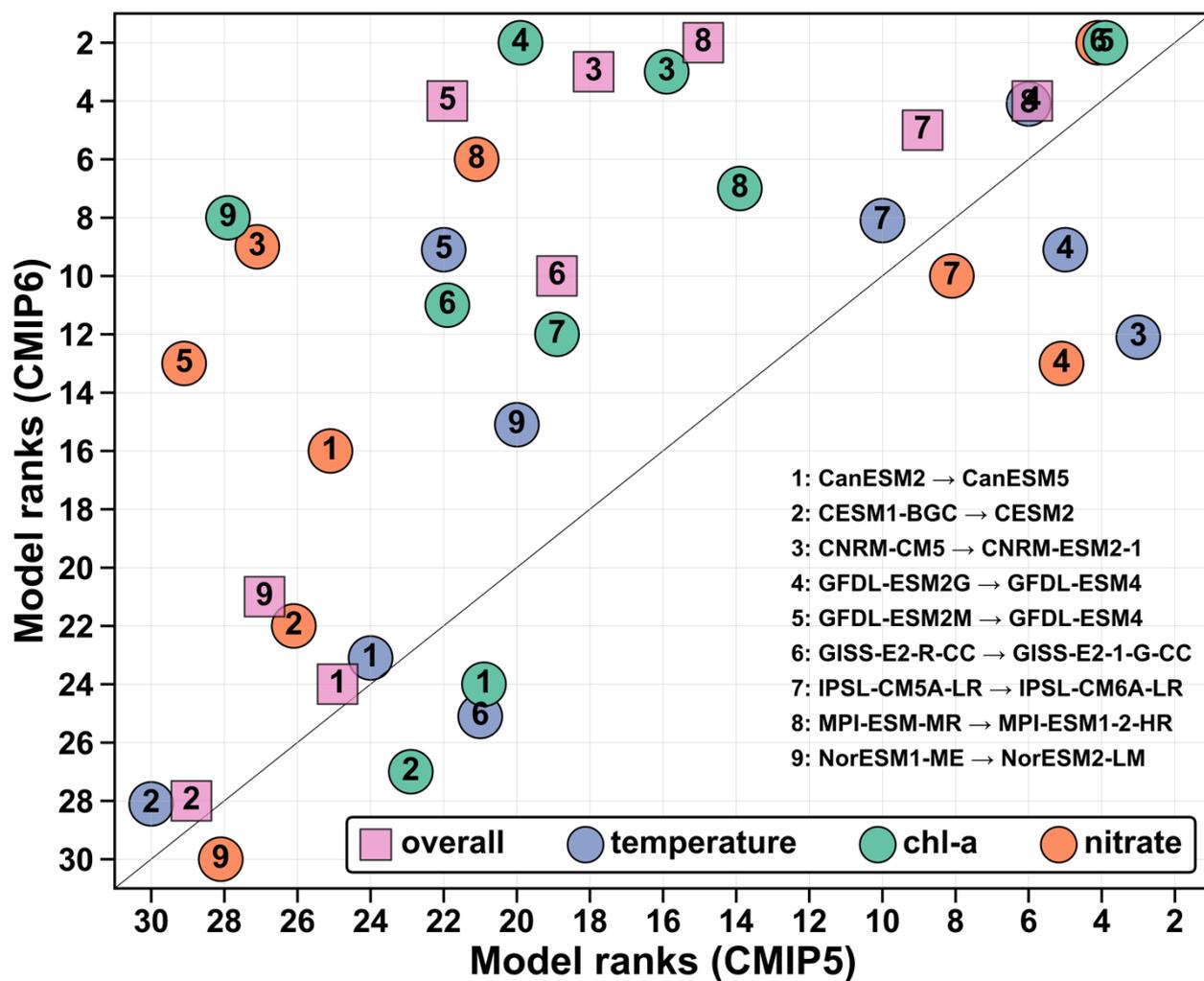
Figure 12. Comparison of the ranks of the former (x-axis) and current (y-axis) generations of ESMs. Blue dots indicate the overall ranks, the green dots the chlorophyll ranks and the orange dots the nitrate ranks. The numbers indicate the model (see legend). These numbers do not correspond to the original model IDs indicated in Table 1. The black line is the 1:1 line. Dots above this line indicate an improvement and dots below the line a worsening of the rank. Note that there were two CMIP5 GFDL models but only one in the CMIP6 group (model pairs 4 and 5).