JGR Oceans

RESEARCH ARTICLE
10.1029/2022JC018965

Key Points:

- The International Ocean Model Benchmarking tool is used to evaluate the performance of Coupled Model Intercomparison Project (CMIP)-class earth system models for upper ocean physical and biogeochemical variables
- Model performance improves overall from CMIP5 to CMIP6, particularly for surface nutrients, surface salinity, and mixed layer depth
- A low bias in model anthropogenic carbon uptake from 1994 to 2007 is related to the vertical temperature gradient between 200 and 1,000 m

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

W. Fu,
weiweif@uci.edu

Citation:

Fu, W., Moore, J. K., Primeau, F., Collier, N., Ogunro, O. O., Hoffman, F. M., & Randerson, J. T. (2022). Evaluation of ocean biogeochemistry and carbon cycling in CMIP earth system models with the International Ocean Model Benchmarking (IOMB) software system. *Journal of Geophysical Research: Oceans*, *127*, e2022JC018965. https://doi.org/10.1029/2022JC018965

Received 10 JUN 2022
Accepted 20 SEP 2022

Evaluation of Ocean Biogeochemistry and Carbon Cycling in CMIP Earth System Models With the International Ocean Model Benchmarking (IOMB) Software System

Weiwei Fu[1], J. Keith Moore[1], Francois Primeau[1], Nathan Collier[2], Oluwaseun O. Ogunro[2], Forrest M. Hoffman[2,3], and James T. Randerson[1]

[1]Department of Earth System Science, University of California, Irvine, CA, USA, [2]Oak Ridge National Laboratory, Climate Change Science Institute (CCSI), Oak Ridge, TN, USA, [3]Department of Civil & Environmental Engineering, University of Tennessee, Knoxville, TN, USA

Abstract The International Ocean Model Benchmarking (IOMB) software package is a new community resource that we use here to evaluate surface and upper ocean biogeochemical variables and integrated anthropogenic carbon uptake from earth system models (ESMs) contributing to the 5th and 6th phases of the Coupled Model Intercomparison Project (CMIP5 and CMIP6). IOMB generates graphics and tables for systematically comparing model predictions against multiple datasets. Our analysis reveals some improvement in the multi-model mean from CMIP5 to CMIP6 for most of the variables we examined. Compared to data-constrained estimates of ocean anthropogenic carbon uptake for the 1994–2007 period, negative biases exist for many models between 30 and 50°S. Global model estimates of anthropogenic carbon uptake for the same period do not change significantly from CMIP5 to CMIP6, with the combined ensemble mean estimate of 27.8 ± 0.5 Pg C lower than a data-constrained estimate of 33.0 ± 4.0 Pg C. At the same time, the change in the natural carbon inventory from CMIP is estimated to be a source of 0.7 ± 0.3 Pg C, which is considerably smaller in magnitude than a data-constrained estimate of 5.0 ± 3.0 Pg C. With chlorofluorocarbon (CFC) predictions available for several models, we demonstrate that negative anthropogenic dissolved inorganic carbon biases coincide with negative biases in CFC concentration, highlighting the importance of weak exchange between the surface and interior ocean in regulating rates of anthropogenic carbon uptake. To examine the robustness of this attribution across the CMIP models, we calculate the global vertical temperature gradient between 200 and 1,000 m as a metric for global stratification and exchange between the surface and deeper waters. We find a linear relationship between the bias of the vertical temperature gradients and the bias in global anthropogenic carbon uptake, consistent with the hypothesis that model biases in anthropogenic carbon uptake are related to biases in surface-to-interior exchange by physical processes.

Plain Language Summary With increasing complexity of earth system models and a rapidly expanding set of ocean observations, we develop an International Ocean Model Benchmarking system, which quantitatively assesses model performance and provides different ways to visualize model-data differences. Our analysis reveals general improvement in the newer generation of ocean biogeochemistry models used to support the 6th Intergovernmental Panel on Climate Change (IPCC) Assessment, compared to an earlier generation of models used to support the 5th IPCC Assessment. A common feature of both generations of ocean models is on average, they absorb less human-emitted carbon dioxide from the atmosphere during a period when observations are available between 1994 and 2007. Comparison of chlorofluorocarbon simulations reported by a few of the models with observations indicates that ocean transport and mixing may be responsible for some of the error in anthropogenic carbon. Comparison with the global temperature profile suggests that in some models, carbon dioxide absorbed at the surface is not pumped into the deeper ocean at a fast-enough rate by circulation and mixing.

1. Introduction

Ocean biogeochemical models are powerful tools to study marine ecosystems, biogeochemistry, and the ocean carbon cycle. Many earth system models (ESMs) have integrated ocean biogeochemical models as essential components of the carbon cycle. At present, the sixth phase of the Coupled Model Intercomparison Project (CMIP6) provides the science community with new ESM projections that draw upon extensive improvements

and incorporation of new features in ocean model components developed by different modeling centers (Eyring et al., 2016; Gidden et al., 2019; Griffies et al., 2016; O'Neill et al., 2016; Orr et al., 2017). Compared with the fifth phase of the CMIP (CMIP5), marine biogeochemical models in CMIP6 have evolved toward more comprehensive representations of plankton communities, sediment biogeochemistry, trace gases relevant to atmospheric chemistry, and variable plankton elemental stoichiometry (Seferian et al., 2020). CMIP6 has provided new projections of ocean acidification, deoxygenation, nutrient distributions, and primary production for scenarios of historical and future change similar to those investigated for CMIP5 (Kwiatkowski et al., 2020). Therefore, a detailed evaluation of model performance and quantification of model biases in comparison with observational datasets for CMIP5 and CMIP6 models is of critical importance for understanding how the models are evolving through time, and whether the increases in model complexity for CMIP6 have yielded commensurate increases in model skill. The knowledge of model performance derived from this assessment can inform studies investigating ocean biogeochemistry, climate variability, future projections, and related downscaling and impact analysis (Flato et al., 2013). In addition, for CMIP6, comprehensive model validation can also support the IPCC Sixth Assessment Report (AR6) (Canadell et al., 2021) and other subsequent climate assessments reports at national and regional scales.

Ocean biogeochemical models within ESMs have increasingly been used for research on the carbon cycle and the transient climate response to cumulative CO_2 emissions in recent decades. A bias in carbon uptake by the oceans may lead to a bias in atmospheric CO_2 concentration in emissions-forced simulations, and consequently biases in the radiative forcing changes due to anthropogenic emissions (Hoffman et al., 2014). For CMIP5, a comparison between the observed and modeled cumulative anthropogenic CO_2 uptake during the historical period was investigated by Bronselaer et al. (2017). The analysis suggested that 6 models had negative biases and 6 models had positive biases, while two models agreed well with observations for the period of 1791–1995, and demonstrated the need to account for changes in the ocean inventory caused by anthropogenic forcing of atmospheric CO_2 prior to the start of CMIP5 simulations in 1850. Additional aspects of ocean carbon uptake were explored in a variety of studies linking carbon cycling with ocean ventilation and circulation (DeVries et al., 2017; Fletcher, 2017; Frolicher et al., 2015), carbon pumps (Yamamoto et al., 2018), and air-sea CO_2 exchange (Dong et al., 2016; Lauderdale et al., 2016). New estimates of anthropogenic carbon uptake derived from repeat ocean transects during 1994–2007 (Gruber et al., 2019) provide a new opportunity to evaluate anthropogenic CO_2 uptake and ocean carbon cycling in the ESMs. The new observations allow for an assessment of the models for a shorter contemporary period when atmospheric carbon dioxide levels were rapidly changing. This new data constraint allows for evaluation of a different set of mechanisms regulating carbon uptake in the ESMs compared to the full anthropogenic inventory change from the pre-industrial period to the present (Sabine et al., 2004; Sabine & Tanhua, 2010).

Here we conduct a model evaluation for several important biogeochemical and physical climate-related variables for both CMIP5 and CMIP6 models using IOMB (the International Ocean Model Benchmarking [IOMB] software package) (Collier et al., 2018; Ogunro et al., 2018). The IOMB package can make quantitative comparisons between time-dependent sequences of observed and simulated multidimensional fields of ocean biogeochemistry data, including sparsely distributed ocean interior observations. We seek to assess whether the representation of ocean biogeochemistry has improved from CMIP5 to the CMIP6. Our second goal is to quantify model biases of anthropogenic ocean carbon uptake in recent decades, and to investigate the relationship between these biases and physical processes in the models.

2. Methods

2.1. CMIP5 and CMIP6 Models

We compare biogeochemical variables from 11 CMIP5 and 9 CMIP6 ESMs (Table 1) with observation-based estimates over the same time period (Table 2), using the IOMB software system. Table 1 provides a summary of the ocean and marine biogeochemical components of these models. These ESMs differ in their horizontal and vertical ocean model resolution, representation of ocean ecosystems and biogeochemistry, as well as the physical processes regulating ocean mixing and overturning. Most of the models revised their representation of marine biogeochemistry from CMIP5 to CMIP6 (Table 1), and entirely new models were incorporated at some climate centers (i.e., the GFDL CM4 model). To evaluate model performance, we use 19th and 20th century simulations of climate change (the CMIP6 "historical" simulation) and a corresponding pre-industrial control simulation

Table 1
The Fifth Phase of the Coupled Model Intercomparison Project (CMIP5) and Sixth Phase of the Coupled Model Intercomparison Project (CMIP6) Ocean Models Evaluated Here Using International Ocean Model Benchmarking

CMIP5			CMIP6		
ESM	Ocean	Ocen BGC	ESM	Ocean	Ocean BGC
MPI-ESM-LR (Giorgetta et al., 2013)	MPI-OM (1°′1.4°)	HAMOCC v5.2 (Ilyina et al., 2013)	MPI-ESM1-2-LR (Muller et al., 2018)	MPI-OM (1.5°′1.5°)	HAMOCC6 (Paulsen et al., 2017)
MPI-ESM-MR (Giorgetta et al., 2013)	MPI-OM (1.41°′0.89°)	HAMOCC v5.2 (Ilyina et al., 2013)	MPI-ESM1-2-HR (Muller et al., 2018)	MPI-OM (0.4°′0.4°)	HAMOCC6 (Paulsen et al., 2017)
IPSL-CM5A-LR (Dufresne et al., 2013)	NEMO-ORCA2 (2°′2°)	PISCES (Aumont & Bopp, 2006)	IPSL-CM6A-LR (Boucher et al., 2020)	NEMO-eORCA1 (1°′1-1/3°)	PISCES v2 (Aumont et al., 2015)
HadGEM2-ES (Collins et al., 2011; Jones et al., 2011)	HadGOM2 (0.3–1°′1°)	Diat-HadOCC (Totterdell, 2019)	UKESM1 (Sellar et al., 2019)	NEMO-ORCA2 (2°′2°)	MEDUSA2 (Yool et al., 2013)
CESM1(BGC) (Gent et al., 2011; Hurrell et al., 2013)	POP2 (1°′1°)	BEC (Moore et al., 2004, 2013)	CESM2 (Danabasoglu et al., 2020)	POP2 (1°′1°)	BEC (Moore et al., 2004, 2013)
NorESM1-ME (Bentsen et al., 2013)	MICOM (1.125°)	HAMOCCv5.1 (Tjiputra et al., 2013)	NorESM2 (Seland et al., 2020)	MICOM-Tripolar (0.5°′0.9°)	iHAMOCC (Tjiputra et al., 2020)
CNRM-CM5 (Voldoire et al., 2013)	NEMO-ORCA1 (1°′1°)	PISCES (Aumont et al., 2003)	CNRM-ESM2-1 (Seferian et al., 2019)	NEMO-eORCA1 (1°′1°)	PISCES v2 (Aumont et al., 2015)
CanESM2 (Arora et al., 2013)	CanOM4 (0.9°′1.4°)	CMOC (Christian et al., 2010; Zahariev et al., 2008)	CanESM5 (Swart et al., 2019)	NEMO-ORCA1 (1°′1-1/3°)	CMOC (Christian et al., 2010; Zahariev et al., 2008)
GFDL-ESM2G (Dunne et al., 2012, 2013)	isopycnal based using GOLD Tripolar (1°′1°)	TOPAZ2 (Dunne et al., 2013)	GFDL-ESM4 (Dunne, Horowitz, et al., 2020)	MOM6 (0.5°)	COBALTv2 (Stock et al., 2014)
GFDL-ESM2M (Dunne et al., 2012, 2013)	MOM4-Tripolar (1°′1°)	TOPAZ2 (Dunne et al., 2013)	GFDL-CM4 (Held et al., 2019)	MOM6 (0.25°)	BLINGv2 (Dunne, Bociu, et al., 2020)

(referred to as the CMIP6 "piControl"). Note that the CMIP6 historical run spans the period from 1850 to 2014 while the CMIP5 historical simulation ends in 2005, with some models starting in 1860 (e.g., HadGEM2-ES) or in 1861 (e.g., GFDL-ESM2M). For the CMIP5 models, we used the RCP8.5 scenario output to extend the time series through 2007.

In the context of this comparison, it is important to note that the CMIP5 and CMIP6 model historical simulations begin from an assumed steady state in the year 1850 and thus miss ocean carbon accumulation associated with the legacy of pre-1850 anthropogenic CO_2 increases. Anthropogenic net carbon emissions continue to have an impact on the air-sea CO_2 flux for decades to centuries after they are emitted. Therefore, we adjust the models to account for the pre-1850 atmospheric CO_2 forcing by adding 0.7 Pg C to their C_{ant} inventories. The 0.7 Pg C adjustment corresponds to the impact of pre-1850 carbon emissions that are still being felt by the ocean during the analysis period of 1994–2007. This value is derived from the impulse response functions used in Bronselaer et al. (2017).

The protocol for the historical simulation includes forcing by a common set of anthropogenic and natural driver variables derived from observations (Eyring et al., 2016). Both natural (e.g., solar variability and volcanic aerosols) and anthropogenic (e.g., greenhouse gas mole fractions, aerosols, and land use) forcing influence climate variability and long-term trends in this simulation. The historical simulation protocol provides an effective means to compare model estimates with observations and to benchmark changes in model performance as individual models evolve over time. Differences in spin-up protocols are known to account for a substantial component of model disparities for biogeochemical fields (e.g., alkalinity, dissolved inorganic carbon), contributing to a relationship between spin-up duration and assessment metrics in the CMIP5 models (Séférian et al., 2016). To account for the impacts of uneven spin up and model drift, we compute the difference between the historical and piControl simulations to estimate cumulative carbon uptake by the oceans. This is a simple way to detrend and

Table 2
Data Products for Different Biogeochemical and Physical Variables Integrated Within International Ocean Model Benchmarking

Observation	Model variable	Data source and references	Temporal coverage
Chlorophyll	chl	GLODAPv2 (Key et al., 2015; Olsen et al., 2016)	1970–2010
		SeaWIFS (Hu et al., 2012; NASA Goddard Space Flight Center et al., 2018)	1997–2010
		MODIS-Aqua (NASA Goddard Space Flight Center et al., 2018)	1997–2010
Oxygen	o2	GLODAPv2 (Key et al., 2015; Olsen et al., 2016)	1970–2010
		WOA2018 (Garcia, Weathers, et al., 2019)	1955–2010
Nitrate	no3	GLODAPv2 (Key et al., 2015; Olsen et al., 2016)	1970–2010
		WOA2018 (Garcia, Locarnini, et al., 2019)	1955–2010
Phosphate	po4	GLODAPv2 (Key et al., 2015; Olsen et al., 2016)	1970–2010
		WOA2018 (Garcia, Locarnini, et al., 2019)	1955–2010
Silicate	si	GLODAPv2 (Key et al., 2015; Olsen et al., 2016)	1970–2010
		WOA2018 (Garcia, Locarnini, et al., 2019)	1955–2010
Dissolved Inorganic Carbon	dissic	GLODAPv2 (Key et al., 2015; Olsen et al., 2016)	1970–2010
		OCIM (DeVries, 2014)	1994–2007
		Gruber (Gruber et al., 2019)	1994–2007
Alkalinity	alk	GLODAPv2 (Key et al., 2015; Olsen et al., 2016)	1970–2010
Mixed layer depth	mlotst	de Boyer Montégut (2004)	1941–2002
Temperature	thetao	GLODAPv2 (Key et al., 2015; Olsen et al., 2016)	1970–2010
		WOA2018 (Locarnini et al., 2019)	1955–2010
		LDEO (Reynolds et al., 2002)	1955–2010
Salinity	so	GLODAPv2 (Key et al., 2015; Olsen et al., 2016)	1970–2010
		WOA2018 (Zweng et al., 2019)	1955–2010

reduce the impacts of varying degrees of ocean spin up across the models. Anthropogenic carbon uptake is defined as the difference between total ocean dissolved inorganic carbon (DIC) and natural DIC (Orr et al., 2017). To keep track of the changes in the natural carbon pool, we use the dissicnat tracer that is computed using a fixed pre-industrial atmospheric CO_2 mole fraction but otherwise responds to the same initial conditions and forcing as the regular DIC tracer (dissic). Thus, to compute the anthropogenic carbon uptake we compute the difference between the total and the natural DIC (i.e., dissic—dissicnat).

For models that do not report a dissicnat tracer, we estimate anthropogenic carbon as the differences between the ocean DIC of historical and pre-industrial control simulations, and correct for the non-steady state natural carbon flux driven by variations in climate and ocean circulation with a Bayesian hierarchical model.

2.2. Bayesian Hierarchical Model

The Bayesian hierarchical model separates the change in the total DIC into contributions from the anthropogenic uptake, the loss of natural carbon, and fluctuations due to internal climate variability, using all available ensembles from each modeling center in the analysis. The statistical model includes model-specific random effects as well as random effects for individual ensemble members that capture the different phases of the internal modes of climate variability in each ensemble member. Thus, the anthropogenic and natural carbon changes can be expressed as follows,

$$\Delta C_{i,j}^{\text{ant}} = \Delta C^{\text{ant}} + \delta_j^{\text{ant}} + \epsilon_{i,j}^{\text{ant}}, \quad (1)$$

$$\Delta C_{i,j}^{\text{nat}} = \Delta C^{\text{nat}} + \delta_j^{\text{nat}} + \epsilon_{i,j}^{\text{nat}}, \quad (2)$$

where ΔC^{ant} and ΔC^{nat} are multi-model mean changes in the anthropogenic natural carbon inventories, δ_j^{ant} and δ_j^{nat} are random effects that are specific to the jth model, $\epsilon_{i,j}^{\text{ant}}$ and $\epsilon_{i,j}^{\text{nat}}$ are random effects for the ith ensemble

member of the *j*th model. The total change in the DIC inventory for the *i*th ensemble member of the *j*th model is then given by $\Delta C^{\text{ant}}_{i,j} + \Delta C^{\text{nat}}_{i,j}$. We assume that the δ and ϵ are normally distributed with zero mean, that is,

$$\delta^{\text{ant}}_{j} \sim N\left(0, \sigma^2_{\text{ant}}\right), \tag{3}$$

$$\delta^{\text{nat}}_{j} \sim N\left(0, \sigma^2_{\text{nat}}\right), \tag{4}$$

$$\epsilon^{\text{ant}}_{i,j} \sim N\left(0, \tau^2_{\text{ant}}\right), \tag{5}$$

$$\epsilon^{\text{nat}}_{i,j} \sim N\left(0, \tau^2_{\text{nat}}\right), \tag{6}$$

where σ_{ant}, σ_{nat}, τ_{ant}, and τ_{nat} are the standard deviations of the δ and ϵ random effects. Thus, we model the spread across ensemble members as random draws from a normal distribution that is common to all the models. This assumption is manifest in the absence of a *j* subscript on the variances τ^2_{ant} and τ^2_{nat}. While this assumption need not be strictly true, we do not have enough ensemble members from all the models to be able to identify differences in the τ^2_{ant} and τ^2_{nat} across the CMIP models. Note, however, that we do take into account model-specific differences in the mean through the fixed effects δ^{ant}_{j} and δ^{nat}_{j}.

With these assumptions, the Bayesian hierarchical model has six adjustable parameters, $\left[\Delta C^{\text{ant}}, \Delta C^{\text{nat}}, \sigma_{\text{ant}}, \sigma_{\text{nat}}, \tau_{\text{ant}}, \tau_{\text{ant}}\right]$. We estimate these parameters from quantities that can be computed directly from the CMIP5 and CMIP6 model output, that is, from the non-shaded quantities in Table 5. Specifically, we draw a Monte-Carlo sample from the posterior probability distribution for the six adjustable parameters using the probabilistic programming language, Stan (Carpenter et al., 2017), from which we compute the marginal posterior mean and standard deviation of each parameter. Throughout the manuscript, quoted uncertainties relating to carbon inventories should be interpreted as ± 1 standard deviation of the posterior probability distribution for the estimated quantity. The Julia and Stan code used to fit the model is provided on GitHub (https://github.com/fprimeau/Bayesian-Hierarchical-Model-for-the-CMIP5-6-anthropogenic-C-uptake).

2.3. Observational Datasets

We compare model simulations with a variety of observations from the 2018 release of the World Ocean Atlas (WOA18; Boyer et al., 2019 and the Global Ocean Data Analysis Project Version 2 [GLODAPv2]; Olsen et al., 2016; Olsen et al., 2019). We also compare the models with SeaWiFS (Hu et al., 2012; NASA Goddard Space Flight Center et al., 2018) and MODIS-Aqua (NASA Goddard Space Flight Center et al., 2018) chlorophyll products. For the satellite chlorophyll products, data in coastal regions have higher uncertainty than in the open ocean as a consequence of turbidity and the presence of suspended particles. We exclude regions where the depth is less than 350 m for the chlorophyll analysis to minimize the impacts higher uncertainty levels in coastal and shelf waters. Mixed layer depths are evaluated against de Boyer Montégut (2004). Table 2 summarizes key observation classes, model variable names, data products, and time intervals integrated within IOMB. ESM output is often compared with a climatological field from the WOA, but IOMB can also calculate the time-dependent statistics, comparing the underlying measurements against the co-located, co-temporal output from the ocean models (Table 2).

A number of key variables from the models, including chlorophyll, nitrate, phosphate, oxygen, DIC, and alkalinity, are compared with observations at the surface in this configuration of IOMB (archived simulations from CMIP5 only recorded monthly output for these variables at the surface). Temperature and salinity are evaluated for the surface, 200 and 700 m depth levels for both the CMIP5 and CMIP6 models. For ocean anthropogenic carbon inventory, we compare model output with two observation-based estimates during the years 1994–2007, one by Gruber et al. (2019) drawing upon the GLODAPv2 observations (hereafter GR2019) and a second estimate from DeVries (2014) based on inverse modeling using the GLODAPv1 data (hereafter DV2014). The GR2019 uses an extended multiple linear regression approach to separate anthropogenic carbon from natural carbon components and variability in DIC induced by ocean biology. In DV2014, ocean circulation is constrained by assimilating various observations in the inverse model. The DV2014 product spans the 1780 to 2016 period; we use the estimates from 1994 to 2007 to provide an independent assessment of recent changes in the ocean carbon inventory. A key difference between the two products is that GR2019 is affected by the ocean circulation variability over 1994–2007, while the DV2014 approach assumes a steady-state circulation.

2.4. Interpolation

To facilitate the comparison, we bin the GLODAPv2 data into a standard three-dimensional grid, which had a $1° \times 1°$ resolution in the horizontal and 33 levels vertically. The vertical layers in IOMB are identical to WOA standard levels, where layer thickness increases with depth. We perform this binning with a monthly resolution and obtain re-gridded GLODAPv2 data with linear interpolation. For the period of 1970–2010, for example, there are 963704 data points (14.7% of all grid cells have data) for temperature, and 159172 data points (2.3% of all grid cells) for CFC11 in the 0–3,000 m depth range, respectively. To obtain a multi-model mean state, we also interpolate output fields from each model to this same standard grid using linear interpolation in the horizontal and vertical directions.

2.5. IOMB

IOMB is a Python-based open-source, multi-model validation tool that can be used to evaluate the performance of CMIP5 and CMIP6 ocean biogeochemistry models. The IOMB package shares some code with its land model benchmarking counterpart, the International Land Model Benchmarking (ILAMB) package (Collier et al., 2018). The long-term objectives of the IOMB project are: (a) to develop internationally accepted benchmarks for ocean model performance, drawing upon international expertise and collaboration, (b) promote the use of these benchmarks by the international community, and (c) strengthen linkages between model and experimental communities in the design of new model tests and ocean observing programs. As a community diagnostic tool, the IOMB has some features that are found in other tools, such as ESMValTools (Eyring et al., 2020), but it also has several distinctive features that make it effective for the evaluation of ocean models, and for comparing across models and across model versions. The ESMValTool focuses on the evaluation of performance metrics for essential climate parameters in chapters of the Intergovernmental Panel on Climate Change (IPCC) report. IOMB can provide general and useful information about model performance, considering for each variable the spatial pattern of bias and root mean square error (RMSE), annual cycle phasing, as well as amplitude of interannual variability and spatial correlation. In addition, IOMB can simultaneously perform variable-to-variable comparison on multiple models for the same time period, facilitating cross-model analyses. IOMB also allows for the assessment of functional relationships between prognostic variables and one or more forcing variables. Further, IOMB can be customized easily for different applications and can incorporate diagnostic updates and new observational datasets from end-users. IOMB was used previously to benchmark aerosol precursors (Ogunro et al., 2018).

Against a seasonal climatology of monthly observations, we use IOMB to calculate a number of diagnostic metrics including bias, RMSE, annual cycle phasing, magnitude of interannual variability, and spatial correlation, which are described in detail in Collier et al. (2018). IOMB provides model performance scores for each metric and generated a single scalar score for each variable by aggregating scores across metrics and datasets (Figure 1). IOMB has the capability to generate model comparisons at multiple depths. IOMB can generate a top-level, interactive summary page for each variable across models. The online version of Figure 1 is available at https://www.ilamb.org/CMIP5v6/IOMB/dashboard.html. Clicking any square on the summary page for a particular variable allows the user to bring up new global maps comparing that model's mean output with the user-selected observational data set (often a choice between GLODAPv2 and WOA2018) (Figure S1 in Supporting Information S1). Additional plots present global maps of the bias (the difference between model and observations) and RMSE.

The overarching goal of the IOMB (and ILAMB) methodology is to generate a synthesis of model performance relative to a collection of reference data products. IOMB utilizes a scoring system that works on the unit interval, where $s = 0$ reflects a poor model and $s = 1$ a good model. Scores are developed for different aspects of performance (bias, RMSE, annual cycle phasing, interannual variability, and spatial correlation) and merged to form an overall score, which is defined as:

$$S_{\text{overall}} = \frac{S_{\text{bias}} + 2S_{\text{rmse}} + S_{\text{phase}} + S_{\text{iav}} + S_{\text{dist}}}{1 + 2 + 1 + 1 + 1} \tag{7}$$

The definition of individual component scores is described in the following.

The bias score as a function of space is $s_{\text{bias}}(\mathbf{x}) = e^{-\epsilon_{\text{bias}}(\mathbf{x})}$ and we can obtain scalar score by averaging the bias score over space ($\mathbf{X}$). The relative error is obtained by nondimensionalization with the centralized RMS of the reference data

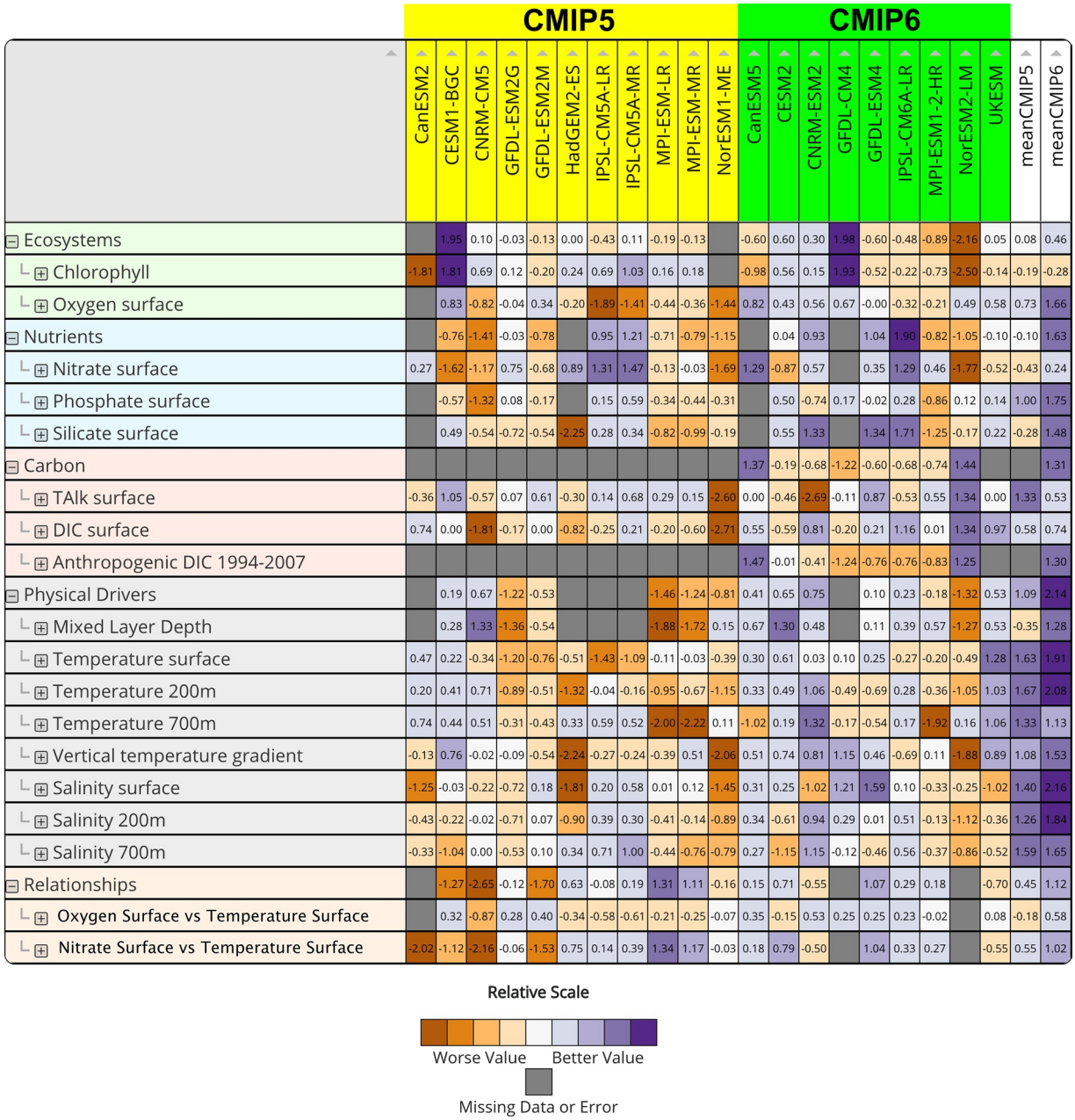

Figure 1. Summary page of the overall score from International Ocean Model Benchmarking (IOMB) for fifth phase of the Coupled Model Intercomparison Project (CMIP5) (yellow background color) and sixth phase of the Coupled Model Intercomparison Project (CMIP6) (green background color) ocean models. In the online version, clicking on any box above brings up additional detailed plots and diagnostics, including the individual skill scores that go into the summary score. The multi-model mean columns for CMIP5 and CMIP6 are constructed by averaging together the maps of each variable from each individual model, and then applying the IOMB package to the resulting mean field.

$$\varepsilon_{\text{bias}}(\mathbf{x}) = |\text{bias}(\mathbf{x})|/\text{crms}(\mathbf{x}) \tag{8}$$

where the bias is defined as $\text{bias}(\mathbf{x}) = \overline{\upsilon_{\text{mod}}(\mathbf{x})} - \overline{\upsilon_{\text{ref}}(\mathbf{x})}$ and the centralized RMS of the reference data is:

$$\text{crms}(\mathbf{x}) = \sqrt{\frac{1}{t_f - t_0}\int_{t_0}^{t_f}\left(\upsilon_{\text{ref}}(t,\mathbf{x}) - \overline{\upsilon_{\text{ref}}}(\mathbf{x})\right)^2 \mathrm{d}t} \tag{9}$$

where t_o is initial time, t_{f} is final time and $\upsilon_{\text{ref}}(t,\mathbf{x})$ is reference data in time and space.

Similarly, the RMSE score is defined as $s_{\mathrm{rmse}}(\mathbf{x}) = e^{-\varepsilon_{\mathrm{rmse}}(\mathbf{x})}$. The relative error of $\varepsilon_{\mathrm{rmse}}(\mathbf{x}) = |\mathrm{rmse}(\mathbf{x})|/\mathrm{crmse}(\mathbf{x})$. We compute the RMSE over the time period of the reference data set with seasonal and interannual variability.

$$\mathrm{rmse}(\mathbf{x}) = \sqrt{\frac{1}{t_{\mathrm{f}} - t_0}\int_{t_0}^{t_{\mathrm{f}}} (v_{\mathrm{mod}}(t,\mathbf{x}) - v_{\mathrm{ref}}(t,\mathbf{x}))^2 \mathrm{d}t} \tag{10}$$

To score the RMSE, we also normalize the centralized RMSE by the centralized RMS of the reference data set.

$$\mathrm{crmse}(\mathbf{x}) = \sqrt{\frac{1}{t_{\mathrm{f}} - t_0}\int_{t_0}^{t_{\mathrm{f}}} \left(\left(v_{\mathrm{mod}}(t,\mathbf{x}) - \overline{v_{\mathrm{mod}}}(\mathbf{x})\right) - \left(v_{\mathrm{ref}}(t,\mathbf{x}) - \overline{v_{\mathrm{ref}}}(\mathbf{x})\right)\right)^2 \mathrm{d}t} \tag{11}$$

The score of phase shift is defined as $s_{\mathrm{phase}}(x) = \frac{1}{2}\left(1 + \cos\left(\frac{2\pi\theta(x)}{365}\right)\right)$. The phase shift of the annual cycle is defined by comparing the timing of the maximum of the annual cycle of the variable, $c(v)$, at each spatial cell across the time period of the reference data set (expressed in days).

$$\theta(\mathbf{x}) = \mathrm{argmax}(c_{\mathrm{mod}}(t,\mathbf{x})) - \mathrm{argmax}(c_{\mathrm{ref}}(t,\mathbf{x})) \tag{12}$$

The score of internal variability is defined as $s_{\mathrm{iav}}(\mathbf{x}) = e^{-\varepsilon_{\mathrm{iav}}(\mathbf{x})}$. The relative error of $\varepsilon_{\mathrm{iav}}(\mathbf{x}) = (\mathrm{iav}_{\mathrm{mod}}(\mathbf{x}) - \mathrm{iav}_{\mathrm{ref}}(\mathbf{x})/\mathrm{iav}_{\mathrm{ref}}(\mathbf{x})$. For the interannual variability, we remove the annual cycle from both the reference and the model.

$$\mathrm{iav}_{\mathrm{mod}}(\mathbf{x}) = \sqrt{\frac{1}{t_f - t_0}\int_{t_0}^{t_f} (v_{\mathrm{mod}}(t,\mathbf{x}) - c_{\mathrm{mod}}(t,\mathbf{x}))^2 \,\mathrm{d}t} \tag{13}$$

The interannual variability in reference data is defined in the same way.

$$\mathrm{iav}_{\mathrm{ref}}(\mathbf{x}) = \sqrt{\frac{1}{t_f - t_0}\int_{t_0}^{t_f} (v_{\mathrm{ref}}(t,\mathbf{x}) - c_{\mathrm{ref}}(t,\mathbf{x}))^2 \,\mathrm{d}t} \tag{14}$$

The score of spatial distribution is defined as $s_{\mathrm{dist}} = \frac{2(1+R)}{\left(\sigma + \frac{1}{\sigma}\right)^2}$, where $\sigma = \frac{\mathrm{stdev}\left(\overline{V_{\mathrm{mod}}}(\mathbf{x})\right)}{\mathrm{stdev}\left(\overline{V_{\mathrm{ref}}}(\mathbf{x})\right)}$ and R is spatial correlation of the period mean value $\overline{V_{\mathrm{mod}}}(\mathbf{x})$ and $\overline{V_{\mathrm{mod}}}(\mathbf{x})$.

More information providing a rational for the individual scores and overall scoring system is described in Collier et al. (2018). In addition, IOMB displays a table summarizing all the statistical metrics that go into the summary page for the selected variable, along with a Taylor diagram (Taylor et al., 2012) showing the overall fit for this variable for all the models (Figures S2 and S3 in Supporting Information S1). Additional plots are generated for some variables relating variability and predictive skill over the annual cycle.

3. Results

3.1. CMIP5 and CMIP6 Model Evaluation With IOMB

Overall, the representation of ocean biogeochemistry improves for most variables from CMIP5 to CMIP6 (Figure 1, Table 3). The scalar scores of model performance are mapped in color, allowing users to visualize the improvement and quickly identify the relative performance of an individual model. The overall scores reported in Figure 1 integrate information on bias and RMSE as well as metrics that quantify differences between the models and the observations for the timing and phase of the annual cycle, interannual variability, and the spatial pattern of the annual mean field (Collier et al., 2018).

To quantitatively assess whether the CMIP6 models show improvement, we report mean estimates from each CMIP using IOMB in two different ways. First, we compute the mean of the scores from individual models (Table 3; first two columns). Second, we create a mean map of each variable, by averaging together maps from individual models within each CMIP. We construct this multi-model mean by interpolating each individual model grid to a $1° \times 1°$ horizontal resolution and to the WOA depth layers in the vertical dimension. We then run IOMB on each of these mean fields (separate mean fields for CMIP5 and CMIP6) and report these scores in the

Table 3
Overall Score of Model Performance for Different Variables in CMIP5 and CMIP6 Models

Variables	Mean of scores from individual CMIP5 models	Mean of scores from individual CMIP6 models	Score derived from the CMIP5 mean field	Score derived from the CMIP6 mean field
Chlorophyll at surface	0.364	0.341	0.365	0.342
Oxygen surface	0.528	0.552	0.560	0.578
Nitrate surface	0.462	0.465	0.455	0.472
Phosphate surface	0.511	0.523	0.538	0.554
Silicate surface	0.421	0.462	0.429	0.500
TALK surface	0.364	0.362	0.394	0.378
DIC surface	0.368	0.392	0.389	0.392
Mixed Layer Depth	0.501	0.562	0.512	0.634
Temperature surface	0.602	0.613	0.641	0.647
Temperature 200 m	0.468	0.481	0.509	0.517
Temperature 700 m	0.463	0.461	0.492	0.487
Salinity surface	0.471	0.483	0.502	0.516
Salinity 200 m	0.468	0.473	0.498	0.510
Salinity 700 m	0.461	0.462	0.498	0.499

Note. The overall score is calculated based on bias, root mean square error, phase shift, interannual variability and spatial distribution. The scoring algorithm is constructed such that higher scores (closer to 1) are better than lower score. The mean of scores from individual models listed in Table 1 are shown in the first two columns, while the application of International Ocean Model Benchmarking to a single mean field constructed from all of the fifth phase of the Coupled Model Intercomparison Project (CMIP5) or sixth phase of the Coupled Model Intercomparison Project (CMIP6) models is shown in the final two columns.

final two columns of Table 3 and in the final two columns of Figure 1. Both approaches reveal that the CMIP6 models have a higher overall score for 10 out of the 13 variables examined. The exceptions where the models do not show improvement include surface chlorophyll, surface ALK, and temperature at 700 m (Figure 1).

To further quantify model improvement, we report the bias and RMSE of the multi-model mean of the CMIP5 and CMIP6 models (Table 4). The bias of multi-model mean is reduced by 20%–70% for surface nitrate, phosphate, and silicate (Table 4) comparing CMIP6 to the CMP5 models. Among them, surface silicate shows the most pronounced improvement. The bias temperature at the surface and at 700 m increases slightly from CMIP5 to CMIP6. For all of the variables we compare in IOMB, RMSE decreases in the CMIP6 models to varying degrees (Table 4). The RMSE of the multi-model mean is generally lower than the mean of individual model RMSE because across-model variability is reduced. In particular, surface DIC and total alkalinity show considerable improvement. The DIC bias is decreased by 42% from 80 to 46 μmol/L and the RMSE is reduced by 46%. Similarly, the total alkalinity bias decreases by 45%, from 79 to 43 μmol/L. The mixed layer depth bias also decreases to −7 m from 16 m in CMIP5, suggesting stronger stratification and perhaps weaker vertical transport in the newer set of models. The full suite of linked graphics and statistical metrics cannot be shown here but is available online (see methods).

While the multi-model means provide evidence for general model improvement, there is no consistent pattern of improvement across models and variables. For example, surface silicate improves in almost all the models, but the improvements are most noteworthy in the HadGEM2 and GFDL-ESM2M newer generation models. Compared with the WOA18 data, the bias is reduced from 45.5 to 2.5 mmol/m^3 from CMIP5 to CMIP6 for HadGEM2, while the bias is reduced from 8.4 to 1.0 mmol/m^3 in GFDL-ESM2G. In CMIP6, the surface silicate of CNRM-ESM2 has the smallest bias of −0.4 mmol/m^3 followed by GFDL-ESM4. The surface silicate of MPI-ESM1-2-HR shows the highest bias of 9.1 mmol/m^3. The improvement in surface silicate suggests CMIP6 models have improved the representation of diatom production and export, which was not included explicitly in some CMIP5 models.

For surface chlorophyll in CMIP5, CESM1-BGC has the highest score followed by the IPSL models, while CanESM2 has the lowest score. However, a relatively high score for CMIP5 does not guarantee a similar ranking for the successor model. In CMIP6, the prediction skill of CESM is degraded for surface chlorophyll while the GFDL-CM4 model is significantly improved (and has the highest score of all the models). The score chart in Figure 1 also calls attention to inconsistent improvements for different variables such as nitrate, phosphate, and silicate within a single model. The model with a high score for one macronutrient may have a poor rating for another, and vice versa. A detailed investigation of these differences is outside the scope of this paper, but further IOMB analysis can help clarify covariances and interactions among some of the drivers.

3.2. Comparison of Anthropogenic DIC Inventory Change With Data-Constrained Estimates

For the CMIP6 models that include the dissicnat variable, we calculate the anthropogenic carbon uptake by computing the difference between the historical DIC and the historical dissicnat inventory change from 1994 to 2007. For these models we can also compute the change in the natural carbon inventory by subtracting the piControl DIC inventory from the historical dissicnat inventory during the same interval. The change in the natural carbon inventory, defined in this way, is a source that ranges in magnitude between 0.5 and 1.1 Pg C for the four models that carry the dissicnat tracer (Table 5). For the other models, we separate anthropogenic and natural carbon inventory change using the Bayesian hierarchical model. We then compare ΔC^{ant} and ΔC^{nat} from the models with data-constrained estimates from DV2014 and GR2019 (Figure 2, Table 5). This comparison is for the integrated carbon storage in the top 3,000 m of the water column.

Table 4
Mean Bias and RMSE of Individual CMIP5 and CMIP6 Models

	Mean of CMIP5		Mean of CMIP6		CMIP5 mean		CMIP6 mean	
Variables	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
Chlorophyll at surface (mg/m³)	0.10	0.44	0.03	0.29	0.09	0.41	0.03	0.24
Oxygen surface (μM)	6.38	12.55	3.27	10.49	6.38	11.21	3.27	8.79
Nitrate surface (μM)	0.84	2.59	0.67	2.29	0.84	1.99	0.67	1.66
Phosphate surface (μM)	−0.08	0.21	−0.02	0.20	−0.12	0.18	−0.06	0.16
Silicate surface (μM)	8.10	10.51	2.21	5.52	8.10	9.21	2.21	4.44
TAlk surface (μmol/L)	78.9	99.7	43.2	64.0	78.9	99.5	43.2	47.8
DIC surface (μmol/L)	80.1	89.9	45.8	57.8	80.12	90.8	45.8	48.3
Mixed Layer Depth (m)	16.1	46.9	−7.4	34.5	16.1	40.5	−7.4	26.4
Temperature surface (°C)	−0.45	1.53	−0.48	1.36	−0.45	1.07	−0.48	1.00
Temperature 200 m (°C)	0.05	1.69	0. 04	1.52	0.05	1.18	0. 04	1.09
Temperature 700 m (°C)	0.49	1.32	0.50	1.29	0.49	1.03	0.51	0.97
Salinity surface (PSU)	−0.23	0.71	−0.08	0.64	−0.23	0.57	−0.08	0.47
Salinity 200 m (PSU)	−0.18	0.37	−0.09	0.34	−0.18	0.27	−0.09	0.22
Salinity 700 m (PSU)	0.03	0.23	0.03	0.22	0.03	0.14	0.03	0.14

Note. The bias and RMSE of multi-model mean variables is also given in the last two columns. The multi-model mean was calculated by interpolating original model grid to 1 × 1° grid in the horizontal. CMIP6, Sixth phase of the Coupled Model Intercomparison Project; CMIP5, fifth phase of the Coupled Model Intercomparison Project; RMSE, root mean square error.

The multi-model mean anthropogenic carbon inventory change (ΔC^{ant}) during the 1994–2007 period is 27.8 ± 2.0 Pg C for CMIP6 and 27.9 ± 1.1 Pg C for CMIP5 (Table 5). For both CMIP5 and CMIP6 models, the multi-model mean anthropogenic and natural carbon changes are estimated to be $\Delta C^{ant} = 27.8 \pm 0.5$ Pg C and $\Delta C^{nat} = 0.7 \pm 0.3$ Pg C. The estimated standard deviations of the model-specific random effects are $\sigma_{ant} = 1.8 \pm 0.4$ Pg C, $\sigma_{nat} = 0.3 \pm 0.4$ Pg C, $\tau_{ant} = 0.2 \pm 0.02$ Pg C, and τ_{nat} =0.4 ± 0.03 Pg C. The multi-model mean is thus smaller than the data-constrained estimates of 30.2 Pg C from DV2014 and 33.0 ± 4.0 Pg C from GR2019. Because DV2014 does not provide an uncertainty estimate, it is difficult to know if the difference is statistically significant. For the case of the GR2019 estimate, which comes with an error estimate, we can compute the cumulative probability distribution for the difference, $\Delta C^{ant}_{CMIP} - \Delta C^{ant}_{GR2019}$, using a Monte-Carlo sampling approach. We accomplish this using the Stan probabilistic programming language. The cumulative probability function, plotted in Figure 3, shows that there is a 90% probability that the CMIP multi-model mean has a negative bias compared to the GR2019 estimate. Section 3.4 discusses some possible reasons for this bias.

The multi-model mean natural carbon loss, is estimated to be $\Delta C^{nat} = -0.8 \pm 0.4$ Pg C for CMIP6 and $\Delta C^{nat} = -0.7 \pm 0.3$Pg C for CMIP5, which is much smaller than the 5.0 ± 3.0 Pg C of natural carbon loss assumed by Gruber et al. (2019). The effect of internal climate variability, as opposed to secular changes in circulation or temperature increases due to global warming, contributes very little to the CMIP6 multi-model mean. This is because the phase of El Niño-Southern Oscillation (ENSO) or Southern Annular Mode (SAM), for example, can be assumed to be a random variable that averages out in the ensemble mean. Even for individual model runs, the effect of internal variability on the change in the inventory of anthropogenic or natural carbon is small. The mean standard deviation for the effect of internal climate variability on the change in anthropogenic carbon is 0.2 Pg C across ensemble members (τ_{ant}; Equation 5), with very little variation in this standard deviation occurring for different CMIP6 models. Similarly, for natural carbon, the impact of internal climate variability yielded a standard deviation of 0.4 Pg C (τ_{nat}; Equation 6) and little variation among individual models. For the change in the anthropogenic carbon inventory, the impact of internal climate variability is more than two orders of magnitude smaller than ΔC^{ant} and is also smaller than the typical across-model differences in the anthropogenic uptake, that is, σ_{ant} =1.8 ± 0.4 Pg C, where 0.4 Pg C is the standard deviation of probability distribution for the estimated σ_{ant}. The relatively small magnitude (and variability) of the global-scale natural climate flux is consistent with previous studies (DeVries et al., 2019; Landschutzer et al., 2016; Schwinger et al., 2014 and references therein).

Table 5
Change of Total, Anthropogenic and Natural DIC Inventory (Pg C) for the 1994 to 2007 Period in the Top 3,000 m of the Water Column for Different CMIP5 and 6 Models

CMIP models and data-constrained estimates		Ensemble size	Total DIC Inventory change (Pg C)	Natural DIC Inventory change (Pg C)	Anthropogenic DIC Inventory change (Pg C)
	DV2014	n.a.	**n.a.**	**n.a.**	30.2
	GR2019	n.a.	28.0 ± 5.0	−5.0 ± 3.0	33.0 ± 4.0
	CMIP mean	157	28.5 ± 0.4	−0.7 ± 0.3	27.8 ± 0.5
	CMIP6 mean	128	28.5 ± 2.1	−0.8 ± 0.4	27.8 ± 2.0
	CMIP5 mean	29	28.6 ± 1.2	−0.7 ± 0.3	27.9 ± 1.1
CMIP6 models	CESM2	9	26.0 ± 0.3	−0.6 ± 0.4	26.6 ± 0.2
	CESM2-WACCM	3	25.8 ± 0.3	−0.5 ± 0.4	26.3 ± 0.2
	CanESM5	25	26.9 ± 0.4	−1.1 ± 0.5	28.0 ± 0.3
	CNRM-ESM2	11	27.4 ± 0.4	−0.8 ± 0.5	28.1 ± 0.5
	GFDL-CM4	1	32.1	−1.1 ± 0.8	33.2 ± 1.1
	GFDL-ESM4	1	29.1	−0.9 ± 0.5	29. 6 ± 0.7
	IPSL-CM6A-LR	35	24.3 ± 0.5	−0.5 ± 0.7	25.2 ± 0.5
	MPI-ESM1-2-HR	10	25.6 ± 0.5	−0.6 ± 0.6	26. 5 ± 0.5
	MPI-ESM1-2-LR	30	26.4 ± 0.6	−0.7 ± 0.5	27.2 ± 0.4
	NorESM2-LM	3	27.2 ± 0.2	−0.7 ± 0.3	27.9 ± 0.1
	CESM1-BGC	1	26.5	−0.8 ± 0.4	27.3 ± 0.6
	CNRM-CM5	1	25.4	−0.8 ± 0.5	26.3 ± 0.6
	CanESM2	5	24.4 ± 0.3	−0.6 ± 0.7	25.4 ± 0.6
	GFDL-ESM2G	1	26.7	−0.8 ± 0.4	27.5 ± 0.6
CMIP5 models	GFDL-ESM2M	1	28.0	−0.7 ± 0.5	28.6 ± 0.7
	HadGEM2-ES	4	26.3 ± 0.4	−0.7 ± 0.5	27.1 ± 0.5
	IPSL-CM5A-LR	3	26.9 ± 0.2	−0.7 ± 0.5	27.6 ± 0.5
	IPSL-CM5A-MR	6	28.8 ± 0.4	−0.9 ± 0.5	29.4 ± 0.6
	MPI-ESM-LR	3	29.0 ± 0.2	−0.9 ± 0.5	29.6 ± 0.7
	MPI-ESM-MR	3	27.0 ± 0.3	−0.7 ± 0.5	27.7 ± 0.5
	NorESM1-ME	1	29.9	−0.6 ± 0.7	30.2 ± 0.8

Note. Also shown are the data-constrained estimates from DV2014 and GR2019. The Arctic and other marginal seas are not included in any of these estimates. The unshaded natural and anthropogenic DIC inventory changes were computed directly from the CMIP6 model output using the "dissicnat" variable. The total DIC inventory change was computed directly from the model output in all cases. The others were estimated from the Bayesian Hierarchical Model. The uncertainties for the shaded numbers denote the standard deviation of the posterior probability distribution. For the non-shaded numbers, the uncertainties denote the standard error. CMIP6, Sixth phase of the Coupled Model Intercomparison Project; CMIP5, fifth phase of the Coupled Model Intercomparison Project; DIC, dissolved inorganic carbon; n.a., not available.

In terms of overall score, CanESM5 and NorESM2-LM models in the CMIP6 have the highest scores for anthropogenic carbon (Figure 1). We note that the comparison of spatial patterns of DIC may increase the uncertainty of the overall score, given the internal variability and non-steady carbon flux driven by climate variability in the CMIP models. However, with the ensembles used, much of this internal variability is expected to average out in the global means.

Most of the models capture some enhanced storage in the Southern Ocean but struggle to reproduce the Southern Hemisphere maximum in storage associated with the formation of Subantarctic Mode and Antarctic Intermediate Waters (Figure 4). The formation of these water masses is key to moving anthropogenic CO_2 from the surface to the ocean interior. The two GFDL models capture the strong storage in mid-latitudes of the Southern Hemisphere, but most models do not, suggesting weak formation of these intermediate and mode waters, at least in some ocean basins.

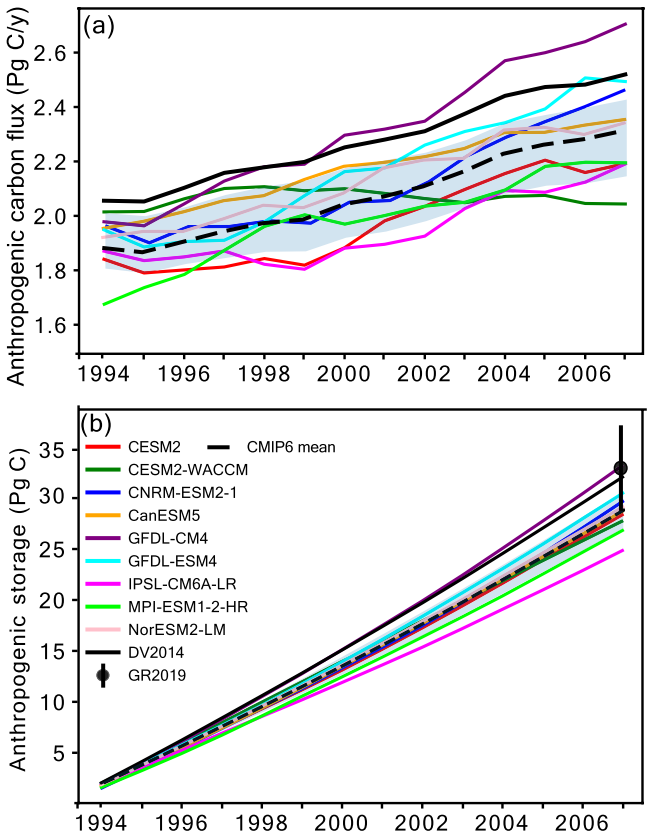

Figure 2. (a) Ocean anthropogenic carbon flux in Pg C/y, and (b) cumulative anthropogenic carbon storage from 1994 to 2007 in the sixth phase of the Coupled Model Intercomparison Project models are compared to the data-constrained estimate from Gruber et al. (2019) and the observation-based inverse estimate from DeVries (2014).

The Southern Ocean south of 40°S accounts for 35% of the global anthropogenic CO_2 uptake from the atmosphere from 1994 to 2007, while storage in the tropics is smaller (Figure 4). The low storage in these regions results from the large transport of anthropogenic carbon out of these regions and higher levels of stratification compared to other areas (Frolicher et al., 2015). The zonal integral of C_{ant} clearly shows the underlying, integrated climate-emissions signals and provides a robust comparison of C_{ant} uptake and distributions (Figure 4).

3.3. Biases of Anthropogenic CO_2, CFCs and Vertical Temperature Gradients

The negative bias in the C_{ant} change for the 1994–2007 period may be attributable to both physical and biogeochemical processes. To assess biases in ocean circulation and mixing, comparison of simulated CFC distributions from CMIP6 models with observations offers the possibility of directly assessing the magnitude of the exchange between surface and sub-surface waters. Because of the known time history of atmospheric concentrations, and the fact that CFCs are biologically and chemically inert in the ocean, they serve as unambiguous tracers of ocean circulation (Dutay et al., 2002; England et al., 1994). By comparison with GLODAPv2 CFC observations, we find that, of the four CMIP6 models that reported CFC values, all had a negative bias in the global ocean inventory over the period of 1994–2007. This suggests that vertical exchange from surface to the interior in the models is too weak. Further analysis reveals that the spatial structure of CFC errors relative to GLODAPv2 and anthropogenic DIC errors relative to GR2019 are positively correlated. The positive correlation is evident in the joint distribution for the DIC and CFC-11 errors (Figure 5). A similar positive correlation is also shown in the joint distribution when the DIC errors are computed relative to DV2014 (Figure S4 in Supporting Information S1). When the DIC error is computed relative to GR2019, the mean of the joint distribution is in the third quadrant indicating that both the DIC and the CFC-11 are negatively biased. The positive correlation suggests that the vertical water-mass exchange is responsible for the biases. For GFDL-CM4 model, the mean CFC error is less negative than for the CESM2 and CESM2-WACCM, consistent with the higher C_{ant} uptake by GFDL-CM4 (Figure 4).

More CFC output from CMIP6 and CMIP5 models is needed to examine the robustness of the relationship between CFC bias and C_{ant} bias. Unfortunately, most modeling centers have not uploaded CFC output to the Earth System Grid (https://esgf-node.llnl.gov/search/cmip6/), even though CFCs are a requested standard output variable for CMIP6. As another tracer of vertical exchange and ocean mixing, we compare the simulated and observed vertical temperature gradient between 200 and 1,000 m. This metric of stratification has several advantages. First, all of the CMIP models report the three-dimensional structure of ocean temperature. Second, thermocline strength is crucial for the ocean carbon sink, and integrates a number of key physical processes. Gnanadesikan (1999) describe the maintenance of the thermocline with a predictive model. Biogeochemical constraints were applied to this model, showing the interacting impacts of vertical and lateral diffusion on thermocline depth (Gnanadesikan et al., 2004). Therefore, the transport of heat and other tracers from the surface to the interior is expected to be related to the strength of the vertical temperature gradients. For the global ocean, the mean vertical temperature gradient may serve as an effective proxy for stratification and vertical exchange.

We compute the temperature gradient (dT/dZ) on each horizontal model grid location using the least squares method. The line of best fit is obtained with vertical temperature profile from 200 to 1000 m as illustrated in Figure S5 in Supporting Information S1. The dT/dZ is similarly computed for the WOA18 and GLODAP v2 data, which is averaged to a $1° \times 1°$ grid. The profiles in Figure S5 in Supporting Information S1 show global mean profiles for the models and the observations. The bias of dT/dZ for different models is shown in Figure S6 in Supporting Information S1 during the period of 1994–2007. The bias of dT/dZ by comparison with the WOA18 data shows similar spatial patterns to the GLODAPv2 data in different basins.

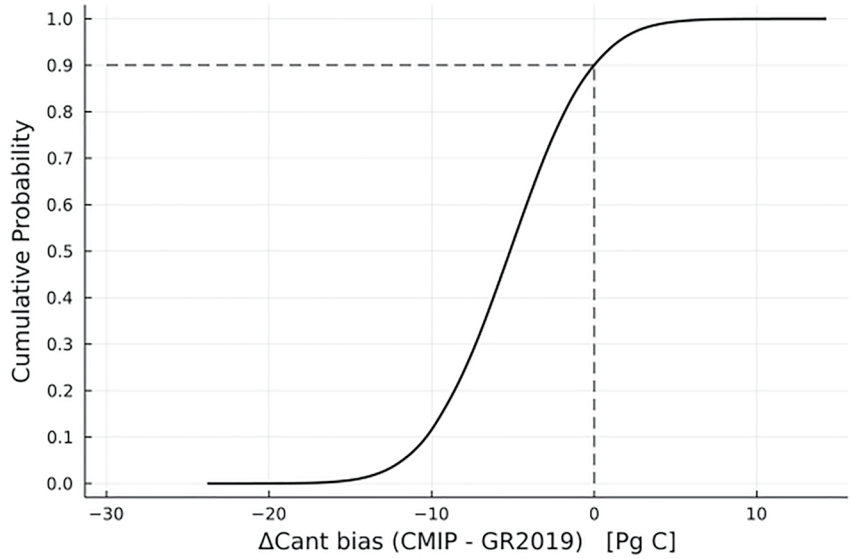

Figure 3. Cumulative probability for the bias in the anthropogenic carbon inventory change in the top 3,000 m of the water column for the 1994–2007 period as estimated from fifth phase of the Coupled Model Intercomparison Project and sixth phase of the Coupled Model Intercomparison Project models listed in Table 5. The bias is computed relative to the data-constrained estimate from GR2019. The graph shows that there is a 90% probability that the Coupled Model Intercomparison Project multi-model mean has a negative bias.

There is a strong correlation between the dT/dZ bias and the bias of DIC inventory across the CMIP5 and the CMIP6 models (Figure 6). We fit a linear function between the bias of C_{ant} inventory and vertical temperature gradient, which has a form of $Y = -2.4\frac{\text{Pg C}}{\text{°C/Km}}\,X - 2.3$ on the global scale. Here, Y is the bias of C_{ant} inventory (Pg C) and X is the bias of the vertical temperature gradient (°C/Km). This negative relationship explains about 60% of model-to-model differences in C_{ant} biases. The weak downward transport, especially in the mid-latitudes of the southern hemisphere, inhibits the transport of anthropogenic CO_2 via the formation of intermediate and Subantarctic mode waters. This is consistent with the large negative bias of C_{ant} storage from 30 to 60°S (Figure 4). Specifically, the IPSL model with the largest positive bias also showed the largest negative C_{ant} bias at these latitudes while GFDL models showed positive C_{ant} biases.

The globally integrated depth-averaged dT/dZ metric shows a strong correlation with the anthropogenic carbon inventory change for the 1997–2004 period. It is important to emphasize that the depth-averaged dT/dZ is not an effective regional metric because it ignores along-isopycnal transport. For instance, the metric did not work as well if restricted in the Southern Ocean (>40°S) even though the Southern Ocean is a major region of carbon uptake. The integrated Southern Ocean DIC inventory is weakly correlated with vertical temperature or vertical density gradients in the Southern Ocean below 40°S (Figure S7 in Supporting Information S1).

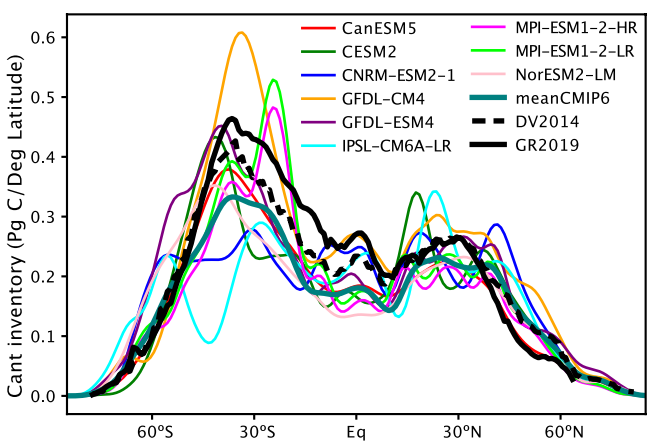

Figure 4. The change of zonally integrated anthropogenic carbon inventory for the period of 1994–2007 as a function of latitude.

4. Discussion and Conclusions

Using the IOMB package, we assess the performance of ocean ecosystem and biogeochemistry models from CMIP5 and CMIP6. We find the performance of CMIP6 models generally better than CMIP5 models, with bias and RMSE reductions for most model variables examined, but the extent of improvement varied depending on variable and individual model. Overall scores improve in the model-mean for CMIP6 for 11 of 14 variables, with exceptions for surface chlorophyll temperature at 700 m and surface alkalinity. These latter variables show small levels of degradation in the multi-model mean. Overall, the summary chart of IOMB is able to provide useful information for future studies of ESMs, and complements the ILAMB package, which evaluates terrestrial physical and biogeochemical variables against observational datasets (Collier et al., 2018).

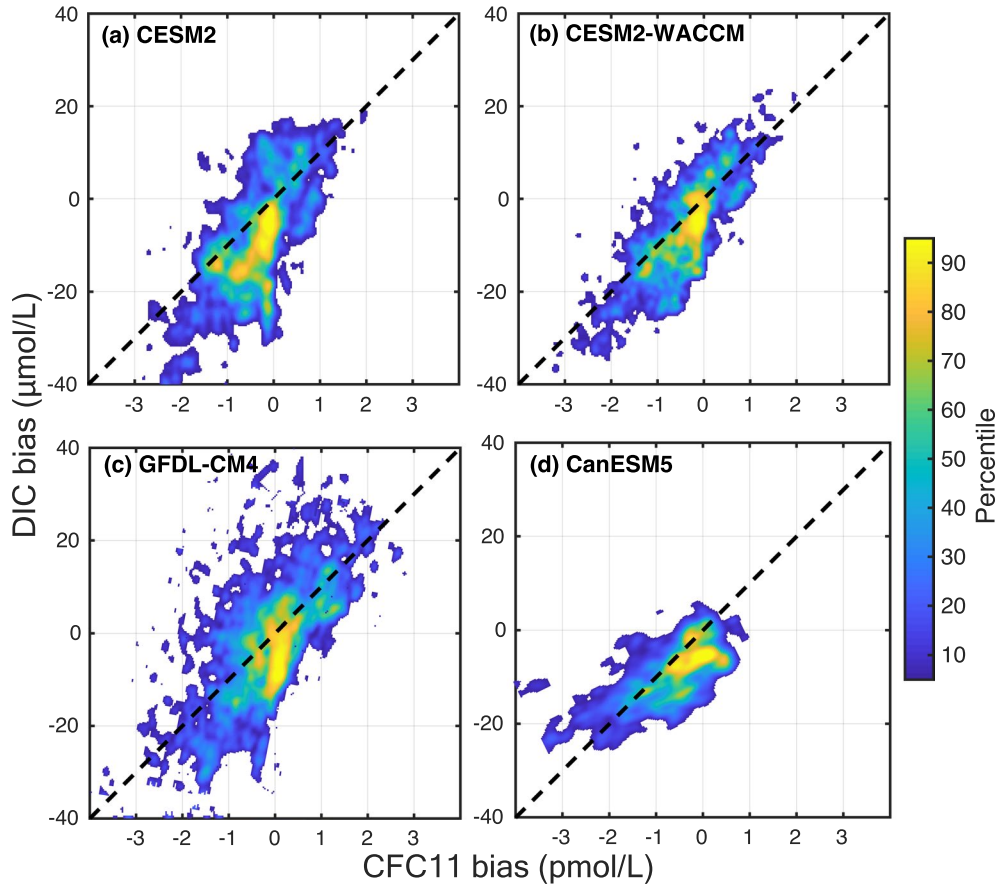

Figure 5. Joint density of anthropogenic ocean dissolved inorganic carbon bias relative to GR2019 and CFC11 bias (relative to GLODAPv2 data) for the period of 1994–2007 for the four models that reported both variables. The bias is calculated by sampling the model where CFC11 observations are available in the 0–3,000 m depth range. There are 83109 points for the period of 1994–2007. The colorbar shows the cumulative density where the *N*th percentile is defined, such that *N*% of the joint distribution lies outside the *N*% shaded colors.

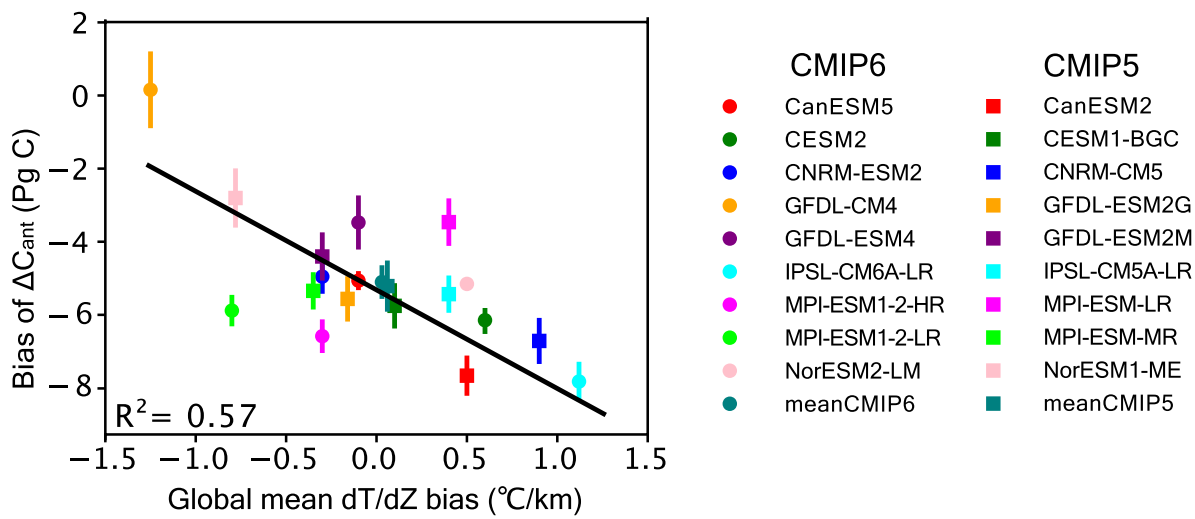

Figure 6. Scatter plot of the global mean bias of anthropogenic ocean dissolved inorganic carbon (DIC) inventory change and the global mean bias of vertical temperature gradient (d*T*/d*Z*) for different fifth phase of the Coupled Model Intercomparison Project (CMIP5) and sixth phase of the Coupled Model Intercomparison Project (CMIP6) models. The bias of anthropogenic DIC inventory for CMIP5 and CMIP6 models is computed relative to the GR2019 for the period of 1994–2007. The uncertainty (ensemble spread) of an individual model indicated by the error bars. The global mean d*T*/d*Z* bias is calculated with the WOA18 data for the period of 1995–2004.

We also estimate the anthropogenic ocean carbon uptake rate and cumulative carbon storage for the historical simulations over the period of 1994–2007. The CMIP6 models predict a multi-model mean of 27.8 ± 2.0 Pg C, which is nearly the same as the 27.9 ± 1.1 Pg C mean from the CMIP5 models. Considering all the models and ensemble members together from CMIP5 and CMIP6, the multi-model mean of 27.8 ± 0.5 Pg C is lower than the estimate of 33.0 ± 4.0 Pg C reported by Gruber et al. (2019). Only the GFDL-CM4 model (one out of 21 CMIP models) has a global mean higher than the data-constrained estimate of ocean anthropogenic carbon storage.

Variability in the ocean's inventory of natural CO_2 emerges from processes such as ocean warming and changes in ocean circulation and biological fluxes in response to climate change (Keeling, 2005; McNeil & Matear, 2013). In Gruber et al. (2019), this non-steady contribution was estimated to be roughly 5.0 ± 3.0 Pg C. However, the CMIP models exhibit a much weaker change in the natural carbon inventory. For the 1994–2007 period the loss of natural carbon in the models that have the dissicnat variable ranges from only 0.5 to 1.1 Pg C. Overall, for the models listed in Table 5 we estimate the multi-model mean loss of natural carbon to be 0.7 ± 0.3 Pg C. Evidence for a relatively small role of a natural carbon cycle response over the time span of a 14-year measurement interval (1994–2007) is also provided in earlier work by Schwinger et al. (2014). Specifically, the mean of 7 CMIP5 models show a cumulative natural carbon response of about 6.7 Pg C per degree of warming over 140 years in an idealized climate experiment (Table 2 of Schwinger et al., 2014). As a back-of-the-envelope calculation, if we assume about 1 K of global surface air temperature warming through 2007, we can divide the Schwinger et al. (2014) estimate by 14 years/140 years, obtaining 0.7 Pg C. This is similar to the estimate we get directly from the CMIP6 models using the natural carbon tracer.

In the CMIP models, ocean internal variability for the period of 1994–2007 may also affect anthropogenic carbon uptake and the change in the natural carbon inventory. We examined the internal variability using all the available ensemble members for each modeling center that provided them. While the carbon inventory can have significant fluctuations at a regional scale, the variability for the globally integrated ocean carbon inventory is small relative to the anthropogenic carbon uptake. The standard deviation for this effect on the anthropogenic carbon inventory change is $\tau_{ant} = 0.2 \pm 0.02$ Pg across initial condition ensembles; for the natural carbon inventory change the standard deviation of this effect is $\tau_{nat} = 0.4 \pm 0.03$. We also expect interannual and decadal climate variability modes to affect changes in anthropogenic and total ocean carbon inventories inferred from the observations (DeVries et al., 2019; McKinley et al., 2020) highlighting the importance of extending the analytical framework developed by Gruber et al. (2019) further in time, so the observed record of anthropogenic carbon change spans multiple decades. This is important for reducing the sensitivity of the observational constraint to climate variability and for gaining insight about the ability of the models to capture carbon cycle processes during a period of time when atmospheric carbon dioxide levels are rapidly changing.

We find a significant relationship between CFC biases and anthropogenic DIC biases in the CMIP models, suggesting vertical exchange is important in structuring some of the low bias in anthropogenic DIC accumulation. The significant negative relationship between the magnitude of the global vertical temperature gradient and anthropogenic carbon uptake provides further evidence that variations in ocean transport and mixing are important for structuring model-to-model differences in their representation of the ocean carbon cycle. More effective evaluation of transport and circulation impacts on anthropogenic carbon uptake in CMIP7 will require more widespread integration and use of CFC and radiocarbon tracers within ocean models; for CMIP6 only three of seven modeling centers reported CFCs.

Comparison of multiple models with a top-level overall scoring chart is an advantage of the IOMB software system over more traditional, single-model, diagnostic tools. It is important to recognize that the single score summarizing model performance is derived using our choice of metrics, which is subjective as in other evaluation tools. In addition to the metrics used in the current version, IOMB will be expanded by incorporating other benchmarking datasets and metrics from the ocean community. The score chart in Figure 1 can be considered an initial, useful evaluation of the CMIP5 and CMIP6 models and can be explored in detail in the online version (see methods).

We evaluated the CMIP6 models at three different depth levels with IOMB. Ongoing IOMB development efforts include adding more model output and observational datasets, adding more types of plots for each variable (comparing with different observational transects, plotting and sampling along isopycnal layers and sub-setting by different ocean basins and/or biomes). The current analysis with the IOMB focuses on a seasonal climatology, but we are adding new features to IOMB for a better assessment of the long-term response of the ocean's carbon

cycle to climate warming. In the long run, we expect IOMB to provide a comprehensive tool for the evaluation of ocean model performance, to help model developers identify deficiencies and subsequently accelerate model development, and to facilitate non-specialist routine analysis for climate and oceanographic studies.

We quantify C_{ant} biases and link them with biases in ocean vertical transport. However, anthropogenic ocean carbon uptake is influenced by many processes, including partial pressure differences, solubility, circulation, and the strength of the biological pump. These processes are related to each other, which compounds the challenge of attribution of the C_{ant} storage bias. The vertical temperature gradient we examined here seems a good metric for vertical exchange in the global ocean. Solubility effects also lead to a positive feedback, which however may be of secondary importance compared with ocean dynamics (Crueger et al., 2008). As shown in Marinov and Gnanadesikan (2011), the storage of ocean carbon is sensitive to ocean circulation, which redistributes the uptake of C_{ant} in the global ocean. In different regions, the relationship between the bias of anthropogenic ocean DIC and vertical temperature gradient requires further exploration with other diagnostics of circulation and mixing. However, the attribution in the global ocean seems robust and the bias in ocean transport appears to be of first order importance in regulating model-to-model differences in the storage of anthropogenic carbon in the oceans. Better representation of the physical processes leading to formation of intermediate and deep water masses should be a high priority for ESM development to improve our ability to project changes in ocean biogeochemistry as climate continues to warm.

Data Availability Statement

The CMIP5 and CMIP6 data can be accessed using the link https://esgf-node.llnl.gov/search/cmip5/ and https://esgf-node.llnl.gov/search/cmip6/. The source code and documentation for IOMB can be found via https://www.ilamb.org/doc/running_iomb.html. The observations for model comparisons and analyses from the IOMB software system are archived in a public repository and can be accessed via the doi: https://doi.org/10.5281/zenodo.6972502.

Acknowledgments
The authors acknowledge support from the Reducing Uncertainty in Biogeochemical Interactions through Synthesis and Computation (RUBISCO) Scientific Focus Area (SFA), which is sponsored by the Regional and Global Model Analysis (RGMA) program area of the Earth and Environmental Systems Sciences Division (EESSD) of the Biological and Environmental Research (BER) office of the U.S. Department of Energy (DOE) Office of Science. FWP acknowledges support from DOE Office of Biological and Environmental Research Award Number DE-SC0021267. We acknowledge the World Climate Research Programme's Working Group on Coupled Modeling, which is responsible for CMIP. We thank the climate modeling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access, and the multiple funding agencies who support CMIP5, CMIP6, and ESGF. We thank DOE's RGMA program area, the Data Management Program, and NERSC for making this coordinated CMIP6 analysis activity possible.

References

Arora, V. K., Boer, G. J., Friedlingstein, P., Eby, M., Jones, C. D., Christian, J. R., et al. (2013). Carbon-concentration and carbon-climate feedbacks in CMIP5 Earth system models. *Journal of Climate*, *26*(15), 5289–5314. https://doi.org/10.1175/Jcli-D-12-00494.1

Aumont, O., & Bopp, L. (2006). Globalizing results from ocean in situ iron fertilization studies. *Global Biogeochemical Cycles*, *20*(2), GB2017. https://doi.org/10.1029/2005gb002591

Aumont, O., Ethe, C., Tagliabue, A., Bopp, L., & Gehlen, M. (2015). PISCES-v2: An ocean biogeochemical model for carbon and ecosystem studies. *Geoscientific Model Development*, *8*(8), 2465–2513. https://doi.org/10.5194/gmd-8-2465-2015

Aumont, O., Maier-Reimer, E., Blain, S., & Monfray, P. (2003). An ecosystem model of the global ocean including Fe, Si, P colimitations. *Global Biogeochemical Cycles*, *17*(2), 1060. https://doi.org/10.1029/2001gb001745

Bentsen, M., Bethke, I., Debernard, J. B., Iversen, T., Kirkevag, A., Seland, O., et al. (2013). The Norwegian Earth System Model, NorESM1-M—Part 1: Description and basic evaluation of the physical climate. *Geoscientific Model Development*, *6*(3), 687–720. https://doi.org/10.5194/gmd-6-687-2013

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., et al. (2020). Presentation and evaluation of the IPSL-CM6A-LR climate model. *Journal of Advances in Modeling Earth Systems*, *12*(7), e2019MS002010. https://doi.org/10.1029/2019MS002010

Boyer, T. P., Antonov, J. I., Baranova, O. K., Garcia, H. E., Johnson, D. R., Mishonov, A. V., et al. (2019). World Ocean database 2018. In A. Mishonov (Ed.), *NOAA Atlas NESDIS* (Vol. *87*).

Bronselaer, B., Winton, M., Russell, J., Sabine, C. L., & Khatiwala, S. (2017). Agreement of CMIP5 simulated and observed ocean anthropogenic CO_2 uptake. *Geophysical Research Letters*, *44*(24), 12298–12305. https://doi.org/10.1002/2017gl074435

Canadell, J. G., Monteiro, P. M., Costa, M. H., Da Cunha, L. C., Cox, P. M., Alexey, V., et al. (2021). Global carbon and other biogeochemical cycles and feedbacks. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, et al. (Eds.) *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. (pp. 673–816). Cambridge University Press. https://doi.org/10.1017/9781009157896.007

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., et al. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–29. https://doi.org/10.18637/jss.v076.i01

Christian, J. R., Arora, V. K., Boer, G. J., Curry, C. L., Zahariev, K., Denman, K. L., et al. (2010). The global carbon cycle in the Canadian Earth system model (CanESM1): Preindustrial control simulation. *Journal of Geophysical Research: Biogeosciences*, *115*(G3), G03014. https://doi.org/10.1029/2008jg000920

Collier, N., Hoffman, F. M., Lawrence, D. M., Keppel-Aleks, G., Koven, C. D., Riley, W. J., et al. (2018). The International Land Model Benchmarking (ILAMB) system: Design, theory, and implementation. *Journal of Advances in Modeling Earth Systems*, *10*(11), 2731–2754. https://doi.org/10.1029/2018ms001354

Collins, W. J., Bellouin, N., Doutriaux-Boucher, M., Gedney, N., Halloran, P., Hinton, T., et al. (2011). Development and evaluation of an Earth-System model-HadGEM2. *Geoscientific Model Development*, *4*(4), 1051–1075. https://doi.org/10.5194/gmd-4-1051-2011

Crueger, T., Roeckner, E., Raddatz, T., Schnur, R., & Wetzel, P. (2008). Ocean dynamics determine the response of oceanic CO_2 uptake to climate change. *Climate Dynamics*, *31*(2–3), 151–168. https://doi.org/10.1007/s00382-007-0342-x

Danabasoglu, G., Lamarque, J., Bacmeister, J., Bailey, D. A., DuVivier, A. K., Edwards, J., et al. (2020). The community earth system model version 2 (CESM2). *Journal of Advances in Modeling Earth Systems*, *12*(2), e2019MS001916. https://doi.org/10.1029/2019MS001916

de Boyer Montégut, C. (2004). Mixed layer depth over the global ocean: An examination of profile data and a profile-based climatology. *Journal of Geophysical Research*, *109*(C12), C12003. https://doi.org/10.1029/2004jc002378

DeVries, T. (2014). The oceanic anthropogenic CO_2 sink: Storage, air-sea fluxes, and transports over the industrial era. *Global Biogeochemical Cycles*, *28*(7), 631–647. https://doi.org/10.1002/2013gb004739

DeVries, T., Holzer, M., & Primeau, F. (2017). Recent increase in oceanic carbon uptake driven by weaker upper-ocean overturning. *Nature*, *542*(7640), 215–218. https://doi.org/10.1038/nature21068

DeVries, T., Le Quere, C., Andrews, O., Berthet, S., Hauck, J., Ilyina, T., et al. (2019). Decadal trends in the ocean carbon sink. *Proceedings of the National Academy of Sciences*, *116*(24), 11646–11651. https://doi.org/10.1073/pnas.1900371116

Dong, F., Li, Y. C., Wang, B., Huang, W. Y., Shi, Y. Y., & Dong, W. H. (2016). Global air-sea CO_2 flux in 22 CMIP5 models: Multiyear mean and interannual variability. *Journal of Climate*, *29*(7), 2407–2431. https://doi.org/10.1175/Jcli-D-14-00788.1

Dufresne, J. L., Foujols, M. A., Denvil, S., Caubel, A., Marti, O., Aumont, O., et al. (2013). Climate change projections using the IPSL-CM5 Earth system model: From CMIP3 to CMIP5. *Climate Dynamics*, *40*(9–10), 2123–2165. https://doi.org/10.1007/s00382-012-1636-1

Dunne, J. P., Bociu, I., Bronselaer, B., Guo, H., John, J. G., Krasting, J. P., et al. (2020). Simple global ocean biogeochemistry with light, iron, nutrients and gas version 2 (BLINGv2): Model description and simulation characteristics in GFDL's CM4.0. *Journal of Advances in Modeling Earth Systems*, *12*(10). https://doi.org/10.1029/2019MS002008

Dunne, J. P., Horowitz, L. W., Adcroft, A. J., Ginoux, P., Held, I. M., John, J. G., et al. (2020). The GFDL Earth system model version 4.1 (GFDL-ESM 4.1): Overall coupled model description and simulation characteristics. *Journal of Advances in Modeling Earth Systems*, *12*(11), e2019MS002015. https://doi.org/10.1029/2019ms002015

Dunne, J. P., John, J. G., Adcroft, A. J., Griffies, S. M., Hallberg, R. W., Shevliakova, E., et al. (2012). GFDL's ESM2 global coupled climate-carbon earth system models. Part I: Physical formulation and baseline simulation characteristics. *Journal of Climate*, *25*(19), 6646–6665. https://doi.org/10.1175/Jcli-D-11-00560.1

Dunne, J. P., John, J. G., Shevliakova, E., Stouffer, R. J., Krasting, J. P., Malyshev, S. L., et al. (2013). GFDL's ESM2 global coupled climate-carbon earth system models. Part II: Carbon system formulation and baseline simulation characteristics. *Journal of Climate*, *26*(7), 2247–2267. https://doi.org/10.1175/Jcli-D-12-00150.1

Dutay, J. C., Bullister, J., Doney, S., Orr, J., Najjar, R., Caldeira, K., et al. (2002). Evaluation of ocean model ventilation with CFC-11: Comparison of 13 global ocean models. *Ocean Modelling*, *4*(2), 89–120. https://doi.org/10.1016/S1463-5003(01)00013-0

England, M. H., Garcon, V., & Minster, J. F. (1994). Chlorofluorocarbon uptake in a world ocean model. 1. Sensitivity to the surface gas forcing. *Journal of Geophysical Research: Oceans*, *99*(C12), 25215–25233. https://doi.org/10.1029/94jc02205

Eyring, V., Bock, L., Lauer, A., Righi, M., Schlund, M., Andela, B., et al. (2020). Earth System Model Evaluation Tool (ESMValTool) v2.0—an extended set of large-scale diagnostics for quasi-operational and comprehensive evaluation of Earth system models in CMIP. *Geoscientific Model Development*, *13*(7), 3383–3438. https://doi.org/10.5194/gmd-13-3383-2020

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the coupled model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., et al. (2013). Evaluation of climate models. In T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Doschung, et al. (Eds.), *Climate change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 741–882). Cambridge University Press. https://doi.org/10.1017/CBO9781107415324.020

Fletcher, S. E. (2017). Climate science: Ocean circulation drove increase in CO_2 uptake. *Nature*, *542*(7640), 169–170. https://doi.org/10.1038/542169a

Frolicher, T. L., Sarmiento, J. L., Paynter, D. J., Dunne, J. P., Krasting, J. P., & Winton, M. (2015). Dominance of the Southern Ocean in anthropogenic carbon and heat uptake in CMIP5 models. *Journal of Climate*, *28*(2), 862–886. https://doi.org/10.1175/Jcli-D-14-00117.1

Garcia, H. E., Locarnini, R. A., Boyer, T. P., Antonov, J. I., Baranova, O. K., Zweng, M. M., et al. (2019). World Ocean Atlas 2018, Volume 4: Dissolved inorganic nutrients (phosphate, nitrate and nitrate + nitrite, silicate). In A. V. Mishonov (Ed.), *NOAA Atlas NESDIS* (Vol. *84*).35.

Garcia, H. E., Weathers, K. W., Paver, C. R., Smolyar, I., Boyer, T. P., Locarnini, M. M., et al. (2019). World Ocean Atlas 2018, Volume 3: Dissolved oxygen, apparent oxygen utilization, and oxygen saturation. In A. V. Mishonov (Ed.), *NOAA Atlas NESDIS* (Vol. *83*).38.

Gent, P. R., Danabasoglu, G., Donner, L. J., Holland, M. M., Hunke, E. C., Jayne, S. R., et al. (2011). The community climate system model version 4. *Journal of Climate*, *24*(19), 4973–4991. https://doi.org/10.1175/2011jcli4083.1

Gidden, M. J., Riahi, K., Smith, S. J., Fujimori, S., Luderer, G., Kriegler, E., et al. (2019). Global emissions pathways under different socioeconomic scenarios for use in CMIP6: A dataset of harmonized emissions trajectories through the end of the century. *Geoscientific Model Development*, *12*(4), 1443–1475. https://doi.org/10.5194/gmd-12-1443-2019

Giorgetta, M. A., Jungclaus, J., Reick, C. H., Legutke, S., Bader, J., Bottinger, M., et al. (2013). Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5. *Journal of Advances in Modeling Earth Systems*, *5*(3), 572–597. https://doi.org/10.1002/jame.20038

Gnanadesikan, A. (1999). A simple predictive model for the structure of the oceanic pycnocline. *Science*, *283*(5410), 2077–2079. https://doi.org/10.1126/science.283.5410.2077

Gnanadesikan, A., Dunne, J. P., Key, R. M., Matsumoto, K., Sarmiento, J. L., Slater, R. D., & Swathi, P. S. (2004). Oceanic ventilation and biogeochemical cycling: Understanding the physical mechanisms that produce realistic distributions of tracers and productivity. *Global Biogeochemical Cycles*, *18*(4), GB4010. https://doi.org/10.1029/2003gb002097

Griffies, S. M., Danabasoglu, G., Durack, P. J., Adcroft, A. J., Balaji, V., Boning, C. W., et al. (2016). OMIP contribution to CMIP6: Experimental and diagnostic protocol for the physical component of the ocean model Intercomparison project. *Geoscientific Model Development*, *9*(9), 3231–3296. https://doi.org/10.5194/gmd-9-3231-2016

Gruber, N., Clement, D., Carter, B. R., Feely, R. A., van Heuven, S., Hoppema, M., et al. (2019). The oceanic sink for anthropogenic CO_2 from 1994 to 2007. *Science*, *363*(6432), 1193–1199. https://doi.org/10.1126/science.aau5153

Held, I. M., Guo, H., Adcroft, A., Dunne, J. P., Horowitz, L. W., Krasting, J., et al. (2019). Structure and performance of GFDL's CM4.0 climate model. *Journal of Advances in Modeling Earth Systems*, *11*(11), 3691–3727. https://doi.org/10.1029/2019ms001829

Hoffman, F. M., Randerson, J. T., Arora, V. K., Bao, Q., Cadule, P., Ji, D., et al. (2014). Causes and implications of persistent atmospheric carbon dioxide biases in Earth System Models. *Journal of Geophysical Research: Biogeosciences*, *119*(2), 141–162. https://doi.org/10.1002/2013jg002381

Hu, C. M., Lee, Z., & Franz, B. (2012). Chlorophyll a algorithms for oligotrophic oceans: A novel approach based on three-band reflectance difference. *Journal of Geophysical Research: Oceans*, *117*(C1), C01011. https://doi.org/10.1029/2011jc007395

Hurrell, J. W., Holland, M. M., Gent, P. R., Ghan, S., Kay, J. E., Kushner, P. J., et al. (2013). The community earth system model a framework for collaborative research. *Bulletin America Meteorology Social*, *94*(9), 1339–1360. https://doi.org/10.1175/Bams-D-12-00121.1

Ilyina, T., Six, K. D., Segschneider, J., Maier-Reimer, E., Li, H. M., & Nunez-Riboni, I. (2013). Global ocean biogeochemistry model HAMOCC: Model architecture and performance as component of the MPI-Earth system model in different CMIP5 experimental realizations. *Journal of Advances in Modeling Earth Systems*, *5*(2), 287–315. https://doi.org/10.1029/2012ms000178

Jones, C. D., Hughes, J. K., Bellouin, N., Hardiman, S. C., Jones, G. S., Knight, J., et al. (2011). The HadGEM2-ES implementation of CMIP5 centennial simulations. *Geoscientific Model Development*, *4*(3), 543–570. https://doi.org/10.5194/gmd-4-543-2011

Keeling, R. F. (2005). Comment on "The ocean sink for anthropogenic CO_2". *Science*, *308*(5729), 1743. https://doi.org/10.1126/science.1109620

Key, R. M., Olsen, A., van Heuven, S., Lauvset, S. K., Velo, A., Lin, X., et al. (2015). *Global Ocean Data Analysis Project, Version 2 (GLODAPv2), ORNL/CDIAC-162, ND-P093*. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, US Department of Energy. https://doi.org/10.3334/CDIAC/OTG.NDP093_GLODAPv2

Kwiatkowski, L., Torres, O., Bopp, L., Aumont, O., Chamberlain, M., Christian, J. R., et al. (2020). Twenty-first century ocean warming, acidification, deoxygenation, and upper-ocean nutrient and primary production decline from CMIP6 model projections. *Biogeosciences*, *17*(13), 3439–3470. https://doi.org/10.5194/bg-17-3439-2020

Landschutzer, P., Gruber, N., & Bakker, D. C. E. (2016). Decadal variations and trends of the global ocean carbon sink. *Global Biogeochemical Cycles*, *30*(10), 1396–1417. https://doi.org/10.1002/2015gb005359

Lauderdale, J. M., Dutkiewicz, S., Williams, R. G., & Follows, M. J. (2016). Quantifying the drivers of ocean-atmosphere CO_2 fluxes. *Global Biogeochemical Cycles*, *30*(7), 983–999. https://doi.org/10.1002/2016gb005400

Locarnini, R. A., Mishonov, A. V., Baranova, O. K., Boyer, T. P., Zweng, M. M., Garcia, H. E., et al. (2019). World Ocean Atlas 2018, Volume 1: Temperature. In A. V. Mishonov (Ed.), *NOAA Atlas NESDIS* (Vol. *81*, 52).

Marinov, I., & Gnanadesikan, A. (2011). Changes in ocean circulation and carbon storage are decoupled from air-sea CO_2 fluxes. *Biogeosciences*, *8*(2), 505–513. https://doi.org/10.5194/bg-8-505-2011

McKinley, G. A., Fay, A. R., Eddebbar, Y. A., Gloege, L., & Lovenduski, N. S. (2020). External forcing explains recent decadal variability of the ocean carbon sink. *AGU Advances*, *1*(2), e2019AV000149. https://doi.org/10.1029/2019AV000149

McNeil, B. I., & Matear, R. J. (2013). The non-steady state oceanic CO_2 signal: Its importance, magnitude and a novel way to detect it. *Biogeosciences*, *10*(4), 2219–2228. https://doi.org/10.5194/bg-10-2219-2013

Moore, J. K., Doney, S. C., & Lindsay, K. (2004). Upper ocean ecosystem dynamics and iron cycling in a global three-dimensional model. *Global Biogeochemical Cycles*, *18*(4), GB4028. https://doi.org/10.1029/2004gb002220

Moore, J. K., Lindsay, K., Doney, S. C., Long, M. C., & Misumi, K. (2013). 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. *Journal of Climate*, *26*(23), 9291–9312. https://doi.org/10.1175/Jcli-D-12-00566.1

Muller, W. A., Jungclaus, J. H., Mauritsen, T., Baehr, J., Bittner, M., Budich, R., et al. (2018). A higher-resolution version of the Max Planck Institute Earth System Model (MPI-ESM1.2-HR). *Journal of Advances in Modeling Earth Systems*, *10*(7), 1383–1413. https://doi.org/10.1029/2017ms001217

NASA Goddard Space Flight Center, Ocean Ecology Laboratory, & O. B. P. Group. (2018). *Sea-viewing Wide Field-of-view Sensor (SeaWiFS) Chlorophyll Data; 2018 Reprocessing (edited)*. NASA OB.DAAC. https://doi.org/10.5067/ORBVIEW-2/SEAWIFS/L3B/CHL/2018

Ogunro, O., Elliott, S., Wingenter, O., Deal, C., Fu, W., Collier, N., & Hoffman, F. (2018). Evaluating uncertainties in marine biogeochemical models: Benchmarking aerosol precursors. *Atmosphere*, *9*(5), 184. https://doi.org/10.3390/atmos9050184

Olsen, A., Key, R. M., van Heuven, S., Lauvset, S. K., Velo, A., Lin, X., et al. (2016). The Global Ocean Data Analysis Project version 2 (GLODAPv2)—An internally consistent data product for the world ocean. *Earth System Science Data*, *8*(2), 297–323. https://doi.org/10.5194/essd-8-297-2016

Olsen, A., Lange, N., Key, R. M., Tanhua, T., Alvarez, M., Becker, S., et al. (2019). GLODAPv2.2019-an update of GLODAPv2. *Earth System Science Data*, *11*(3), 1437–1461. https://doi.org/10.5194/essd-11-1437-2019

O'Neill, B. C., Tebaldi, C., van Vuuren, D. P., Eyring, V., Friedlingstein, P., Hurtt, G., et al. (2016). The scenario model Intercomparison project (ScenarioMIP) for CMIP6. *Geoscientific Model Development*, *9*(9), 3461–3482. https://doi.org/10.5194/gmd-9-3461-2016

Orr, J. C., Najjar, R. G., Aumont, O., Bopp, L., Bullister, J. L., Danabasoglu, G., et al. (2017). Biogeochemical protocols and diagnostics for the CMIP6 ocean model Intercomparison project (OMIP). *Geoscientific Model Development*, *10*(6), 2169–2199. https://doi.org/10.5194/gmd-10-2169-2017

Paulsen, H., Ilyina, T., Six, K. D., & Stemmler, I. (2017). Incorporating a prognostic representation of marine nitrogen fixers into the global ocean biogeochemical model HAMOCC. *Journal of Advances in Modeling Earth Systems*, *9*(1), 438–464. https://doi.org/10.1002/2016ms000737

Reynolds, R. W., Rayner, N. A., Smith, T. M., Stokes, D. C., & Wang, W. Q. (2002). An improved in situ and satellite SST analysis for climate. *Journal of Climate*, *15*(13), 1609–1625. https://doi.org/10.1175/1520-0442(2002)015

Sabine, C. L., Feely, R. A., Gruber, N., Key, R. M., Lee, K., Bullister, J. L., et al. (2004). The oceanic sink for anthropogenic CO_2. *Science*, *305*(5682), 367–371. https://doi.org/10.1126/science.1097403

Sabine, C. L., & Tanhua, T. (2010). Estimation of anthropogenic CO_2 inventories in the ocean. *Annual Review of Marine Science*, *2*(1), 175–198. https://doi.org/10.1146/annurev-marine-120308-080947

Schwinger, J., Tjiputra, J. F., Heinze, C., Bopp, L., Christian, J. R., Gehlen, M., et al. (2014). Nonlinearity of ocean carbon cycle feedbacks in CMIP5 earth system models. *Journal of Climate*, *27*(11), 3869–3888. https://doi.org/10.1175/Jcli-D-13-00452.1

Seferian, R., Berthet, S., Yool, A., Palmieri, J., Bopp, L., Tagliabue, A., et al. (2020). Tracking improvement in simulated marine biogeochemistry between CMIP5 and CMIP6. *Current Climate Change Reports*, *6*(3), 1–25. https://doi.org/10.1007/s40641-020-00160-0

Séférian, R., Gehlen, M., Bopp, L., Resplandy, L., Orr, J. C., Marti, O., et al. (2016). Inconsistent strategies to spin up models in CMIP5: Implications for ocean biogeochemical model performance assessment. *Geoscientific Model Development*, *9*(5), 1827–1851. https://doi.org/10.5194/gmd-9-1827-2016

Seferian, R., Nabat, P., Michou, M., Saint-Martin, D., Voldoire, A., Colin, J., et al. (2019). Evaluation of CNRM Earth system model, CNRM-ESM2-1: Role of earth system processes in present-day and future climate. *Journal of Advances in Modeling Earth Systems*, *11*(12), 4182–4227. https://doi.org/10.1029/2019ms001791

Seland, O., Bentsen, M., Olivie, D., Toniazzo, T., Gjermundsen, A., Graff, L. S., et al. (2020). Overview of the Norwegian Earth System Model (NorESM2) and key climate response of CMIP6 DECK, historical, and scenario simulations. *Geoscientific Model Development*, *13*(12), 6165–6200. https://doi.org/10.5194/gmd-13-6165-2020

Sellar, A. A., Jones, C. G., Mulcahy, J. P., Tang, Y., Yool, A., Wiltshire, A., et al. (2019). UKESM1: Description and evaluation of the U.K. Earth system model. *Journal of Advances in Modeling Earth Systems*, *11*(12), 4513–4558. https://doi.org/10.1029/2019ms001739

Stock, C. A., Dunne, J. P., & John, J. G. (2014). Global-scale carbon and energy flows through the marine planktonic food web: An analysis with a coupled physical-biological model. *Progress in Oceanography*, *120*, 1–28. https://doi.org/10.1016/j.pocean.2013.07.001

Swart, N. C., Cole, J. N. S., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., et al. (2019). The Canadian earth system model version 5 (CanESM5.0.3). *Geoscientific Model Development*, *12*(11), 4823–4873. https://doi.org/10.5194/gmd-12-4823-2019

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of Cmip5 and the experiment design. *Bulletin America Meteorology Social*, *93*(4), 485–498. https://doi.org/10.1175/Bams-D-11-00094.1

Tjiputra, J. F., Roelandt, C., Bentsen, M., Lawrence, D. M., Lorentzen, T., Schwinger, J., et al. (2013). Evaluation of the carbon cycle components in the Norwegian Earth System Model (NorESM). *Geoscientific Model Development*, *6*(2), 301–325. https://doi.org/10.5194/gmd-6-301-2013

Tjiputra, J. F., Schwinger, J., Bentsen, M., Moree, A. L., Gao, S., Bethke, I., et al. (2020). ocean biogeochemistry in the Norwegian Earth System Model Version 2 (NorESM2). *Geoscientific Model Development*, *13*(5), 2393–2431. https://doi.org/10.5194/gmd-13-2393-2020

Totterdell, I. J. (2019). Description and evaluation of the Diat-HadOCC model v1.0: The ocean biogeochemical component of HadGEM2-ES. *Geoscientific Model Development*, *12*(10), 4497–4549. https://doi.org/10.5194/gmd-12-4497-2019

Voldoire, A., Sanchez-Gomez, E., Salas y Melia, D., Decharme, B., Cassou, C., Senesi, S., et al. (2013). The CNRM-CM5.1 global climate model: Description and basic evaluation. *Climate Dynamics*, *40*(9–10), 2091–2121. https://doi.org/10.1007/s00382-011-1259-y

Yamamoto, A., Abe-Ouchi, A., & Yamanaka, Y. (2018). Long-term response of oceanic carbon uptake to global warming via physical and biological pumps. *Biogeosciences*, *15*(13), 4163–4180. https://doi.org/10.5194/bg-15-4163-2018

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

Zahariev, K., Christian, J. R., & Denman, K. L. (2008). Preindustrial, historical, and fertilization simulations using a global ocean carbon model with new parameterizations of iron limitation, calcification, and N-2 fixation. *Progress in Oceanography*, *77*(1), 56–82. https://doi.org/10.1016/j.pocean.2008.01.007

Zweng, M. M., Seidov, D., Boyer, T. P., Locarnini, M., Garcia, H. E., Mishonov, A. V., et al. (2019). World Ocean Atlas 2018, Volume 2: Salinity. In A. Mishonov (Ed.), *NOAA Atlas NESDIS* (Vol. *82*, p. 50).