AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Oceans

RESEARCH ARTICLE

10.1002/2015JC010759

Assessing the abilities of CMIP5 models to represent the seasonal cycle of surface ocean pCO_2

Darren J. Pilcher[1], Sarah R. Brody[2], Leah Johnson[3], and Benjamin Bronselaer[4]

[1]Department of Atmospheric and Oceanic Sciences, University of Wisconsin-Madison, Madison, Wisconsin, USA, [2]Division of Earth and Ocean Sciences, Duke University, Durham, North Carolina, USA, [3]School of Oceanography, University of Washington, Seattle, Washington, USA, [4]Department of Physics, University of Oxford, Oxford, UK

Key Points:
- Model performance highly variable across biomes
- Particular models better suited for specific biomes and metrics
- Internal variability low compared to intermodel variability

Correspondence to:
D. J. Pilcher,
djpilcher@wisc.edu

Citation:
Pilcher, D. J., S. R. Brody, L. Johnson, and B. Bronselaer (2015), Assessing the abilities of CMIP5 models to represent the seasonal cycle of surface ocean pCO_2, *J. Geophys. Res. Oceans*, *120*, 4625–4637, doi:10.1002/2015JC010759.

Received 2 FEB 2015
Accepted 5 JUN 2015
Accepted article online 11 JUN 2015
Published online 3 JUL 2015

Abstract The ability of Earth System Models to accurately simulate the seasonal cycle of the partial pressure of CO_2 in surface water (pCO_2^{SW}) has important implications for projecting future ocean carbon uptake. Here we develop objective model skill score metrics and assess the abilities of 18 CMIP5 models to simulate the seasonal mean, amplitude, and timing of pCO_2^{SW} in biogeographically defined ocean biomes. The models perform well at simulating the monthly timing of the seasonal minimum and maximum of pCO_2^{SW}, but perform somewhat worse at simulating the seasonal mean values, particularly in polar and equatorial regions. The results also illustrate that a single "best" model can be difficult to determine, despite an analysis restricted to the seasonality of a single variable. Nonetheless, groups of models tend to perform better than others, with significant regional differences. This suggests that particular models may be better suited for particular regions, though we find no evidence for model tuning. Timing and amplitude skill scores display a weak positive correlation with observational data density, while the seasonal mean scores display a weak negative correlation. Thus, additional mapped pCO_2^{SW} data may not directly increase model skill scores; however, improved knowledge of the dominant mechanisms may improve model skill. Lastly, we find skill score variability due to internal model variability to be much lower than variability within the CMIP5 intermodel spread, suggesting that mechanistic model differences are primarily responsible for differences in model skill scores.

1. Introduction

Anthropogenic emissions of CO_2 since the Industrial Revolution are a significant perturbation to the global carbon cycle. Of the 555 PgC (1 PgC = 10^{15} gC) emitted due to human industrial activities, approximately half has remained in the atmosphere, increasing atmospheric CO_2 concentrations from 278 ppm in 1750 to 390.5 ppm by 2011 [*Ciais et al.*, 2013]. The remaining emissions have been absorbed into both the terrestrial and ocean carbon sinks. However, the terrestrial biosphere has only acted as a net carbon sink since the 1940s, and is a net source of CO_2 when integrating over the entire industrial period [*Khatiwala et al.*, 2009], leaving the ocean as the only long-term sink over this period. Anthropogenic CO_2 emissions are projected to continue to increase, indicating that the ocean will have to increase CO_2 uptake in order to continue offsetting roughly 25% of human emissions. However, climate-carbon feedbacks associated with ocean warming and changing wind stress are projected to decrease ocean CO_2 uptake [*Friedlingstein et al.*, 2006].

Atmosphere-Ocean Global Climate Models (AOGCMs) are mathematical models that simulate the general circulation of the atmosphere-ocean system. Recent Intergovernmental Panel on Climate Change (IPCC) assessment reports has utilized a developing category of AOGCMs called Earth System Models (ESMs). Although there is not yet a universal definition for an ESM, a fully coupled carbon cycle is often used as a defining characteristic [*Lindsay et al.*, 2014]. ESMs are utilized to determine historical and future ocean carbon uptake and offer spatial and temporal resolution unavailable in observational data. Additionally, they provide experimental earth systems that can be tested under future scenarios of atmospheric greenhouse gas emissions. Projections reported in the IPCC fifth assessment report (IPCC AR5) were produced from the Coupled Model Intercomparison Project phase-5 (CMIP5) [*Taylor et al.*, 2011].

Validating and testing the accuracy of these models is vital toward identifying key processes that are missing and ultimately improving how representative they are of the Earth system. Previous studies tend to

base model skill on the accuracy of hindcast simulations [*Lin*, 2007; *Gleckler et al*., 2008; *Radić and Clarke*, 2011]. Many of these studies find that models that are best in simulating a specific region are often not as skillful in other regions [*Schneider et al*., 2008; *Scherrer*, 2011; *Anav et al*., 2013]. However, these differences are substantially reduced on the global mean scale. For example, models simulate the global ocean CO_2 flux reasonably well, despite much larger model spread in simulating specific regions or variables [*Anav et al*., 2013]. However, model spread increases for projections of global ocean carbon uptake over the 21st Century, due to carbon-climate feedbacks and regional differences in ocean carbon storage and circulation [*Friedlingstein et al*., 2006]. There is also growing interest in differentiating between model inaccuracies due to systematic or process-based errors and errors related to the impact of internal variability [*Deser et al*., 2012a]. The former is indicative of model deficiencies, whereas the latter is the influence of natural variability as simulated within the Earth system. The degree to which internal variability can influence model skill and rankings is an important question that has often not been addressed.

The direction of the air-sea flux of CO_2 is determined by the gradient in the partial pressure of CO_2 (pCO_2) between the atmosphere and the surface ocean. To first order, surface ocean pCO_2 (pCO_2^{SW}) will tend toward equilibrium with the atmosphere. However, pCO_2^{SW} is also a function of temperature, salinity, alkalinity, and dissolved inorganic carbon (DIC). Thus, oceanic physical and biological processes can substantially modify pCO_2^{SW} and generate regions of both net carbon uptake and efflux. Equatorial regions are supersaturated in pCO_2^{SW} due to DIC upwelling and are therefore annual net sources of CO_2 to the atmosphere [*Takahashi et al*., 2009; *Landschützer et al*., 2014a, 2014b]. Conversely, midlatitude to high-latitude regions are generally undersaturated in pCO_2^{SW} due to colder water temperatures and greater biological productivity, and are annual net CO_2 sinks. An exception is the Polar Southern Ocean, where observational data suggest a slight annual source of CO_2, albeit with a relatively sparse observational record [*Takahashi et al*., 2009; *Landschützer et al*., 2014a, 2014b]. Seasonality plays a much greater role in high-latitude regions, with areas such as the North Pacific alternating between a CO_2 source in winter and a CO_2 sink in summer [*Landschützer et al*., 2014a, 2014b]. Significant seasonality combined with limited wintertime observations can bias estimates of ocean anthropogenic carbon uptake [*Rodgers et al*., 2008]. Furthermore, future ocean carbon uptake may become more dependent on the seasonal drawdown of pCO_2 by biological productivity, due to reduced ocean buffer capacity [*Hauck and Völker*, 2015]. Considering, for example, that the MPI models strongly overestimate the high-latitude Southern Hemisphere CO_2 sink in austral summer [*Anav et al*., 2013], this seasonal bias may become amplified at reduced buffer capacity. Thus, differences in the seasonal cycle of pCO_2^{SW} between CMIP5 models may impact projections for the evolution of ocean carbon uptake over the 21st Century, and highlight needed areas of improvement in the physical or biological components of the models.

Here we design and implement model skill metrics to assess the abilities of CMIP5 models to reproduce the seasonal mean, timing, and magnitude of pCO_2^{SW}. We expand upon previous model skill score studies by focusing specifically on the seasonality of a single variable (pCO_2^{SW}), by incorporating biogeographically distinct ocean biomes, and providing a quantitative estimate for the impact of internal variability on our model skill scores.

2. Methods

2.1. CMIP5 Models

We utilized output from 18 global simulations for the recent IPCC Fifth Assessment Report (AR5) as part of CMIP5. The CMIP5 archive includes a wide range of experiments using historical and projected 21st Century forcing. The archived data can be accessed via the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and is freely available to the research community (http://cmip-pcmdi.llnl.gov/cmip5/). Details of these models and their relevant ocean components are listed in Table 1.

We specifically analyze models that contain monthly output of pCO_2^{SW} over the historical period. Within these models, only a subsection of them provided an ESM historical run (refer to Table 1). These data are used to generate a monthly pCO_2^{SW} climatology for each model by averaging over a 10 year window between 1995 and 2005, to match the observational climatology. This monthly climatology is then regridded to the coarser $4° \times 5°$ grid of the observational climatology using an objective analysis interpolation scheme [*Barnes*, 1994].

Table 1. CMIP5 Models Included in This Study and Corresponding Ocean Model, Vertical Coordinate Specification, Biological Model, and Chemical Versus Ecosystem Framework

Full Name	Model Name	Model Type	Ocean Physics	Isopycnal Versus Z-Level	Ocean Biology	Chem Versus Ecosystem
Beijing Climate Center, China Meteorological Administration	BCC-CSM1.1	ESM Historical	MOM (4)	z-level	MOM based on OCMIP2	Chemical
Community Earth System Model Contributors (NFS, DOE, NCAR)	CESM1(BGC)	ESM Historical	POP2	z-level	BEC	Ecosystem
Centro Euro-Mediterraneo per I Cambiamenti Climatici	CMCC-CESM	Historical	NEMO	z-level	PELAGOS/BFM	Ecosystem
Canadian Center for Climate Modelling and Analysis	CanESM2	ESM Historical	OGCM4/CanOM4	z-level	CMOC	Ecosystem (NPZD)
The First Institute of Oceanography, SOA, China	FIO-ESM	ESM Historical	POP2	z-level	OCMIP2	Chemical
NOAA Geophysical Fluid Dynamics Laboratory	GFDL-ESM2M	ESM Historical	MOM	z-level	TOPAZ2	Ecosystem
NOAA Geophysical Fluid Dynamics Laboratory	GFDL-ESM2G	ESM Historical	GOLD	isopycnal	TOPAZ2	Ecosystem
NASA Goddard Institute for Space Studies	GISS-E2-H-CC	Historical	HYCOM	isopycnal/ z-level hybrid		Chemical
NASA Goddard Institute for Space Studies	GISS-E2-R-CC	Historical	Russell ocean Model	z-level		Chemical
Met Office Hadley Centre	HadGEM2-CC	Historical	NEMO	z-level	Diat-HadOCC	Ecosystem
Met Office Hadley Centre	HadGEM2-ES	ESM Historical	NEMO	z-level	Diat-HadOCC	Ecosystem
Institute for Numerical Mathematics	INM-CM4	ESM Historical	INMCM4	z-level		
Japan Agency for Marine-Earth Science and Technology, Atmospheric and Ocean Research Institute (the University of Tokyo) and National Institute for Environmental Studies	MIROC-ESM	ESM Historical	COCO	σ/zlevel hybrid	NPZD	Ecosystem
Japan Agency for Marine-Earth Science and Technology, Atmospheric and Ocean Research Institute (the University of Tokyo) and National Institute for Environmental Studies	MIROC-ESM-CHEM	Historical	COCO	σ/zlevel hybrid	NPZD	Ecosystem
Max-Planck Intitut fur Meteorologie	MPI-ESM-LR	Historical	MPIOM	z-level	HAMOCC5	Ecosystem
Max-Planck Intitut fur Meteorologie	MPI-ESM-MR	Historical	MPIOM	z-level	HAMOCC5	Ecosystem
Meteorological Research Institute	MRI-ESM1	ESM Historical	MRI.COM4	isopycnal/ zlevel hybrid		
Norwegian Climate Centre	NorESM1-ME	ESM Historical	MICOM	isopycnal	HAMOCC	Ecosystem

2.2. Observational pCO_2 Data

Modeled monthly pCO_2^{SW} climatology is compared to the monthly pCO_2^{SW} climatology of *Takahashi et al.*, [2009]. This climatology is constructed from over 3 million pCO_2^{SW} shipboard measurements collected since the 1970s. The climatology has a spatial resolution of 4° latitude × 5° longitude and is referenced to the year 2000. Measurements from coastal regions, as well as those from the Equatorial Pacific collected during El Niño years are excluded [*Takahashi et al.*, 2009].

2.3. Model Biomes

Because pCO_2^{SW} is controlled by biological and physical processes that vary regionally, we use the *Fay and McKinley* [2014] criteria to define biomes consistent with the global biogeography of the ocean. Boundaries of these 17 biomes are derived from sea surface temperature (SST), maximum mixed layer depth, sea ice concentration, and chlorophyll *a* concentrations (Figure 1). We downloaded the *Fay and McKinley* [2014] mean biome definitions (http://doi.pangaea.de/10.1594/PANGAEA.828650) and regridded the definitions to a 4° × 5° grid.

Most (although not all) of the CMIP5 models in our study also output the variables needed to define the *Fay and McKinley* [2014] biomes. We tested the effect of using variable biome definitions (i.e., biomes calculated from the output of a specific CMIP5 model) as opposed to fixed biome definitions (i.e., biomes calculated from observational data). The goal of this experiment was to determine if model skill scores were biased by the specific boundaries of the biomes in each model. We tested a subset of the CMIP5 models that contained monthly SST, chlorophyll, and mixed layer depth output data (GFDL-ESM2G, GFDL-ESM2M, and MRI-ESM1). The results of this test (Figure 2) produced a slight (generally <0.1), albeit discernible difference in model skill score between the fixed and variable biome definitions, but revealed no clear bias. With no evidence for an underlying bias and in the interest of keeping the comparison consistent across the CMIP5

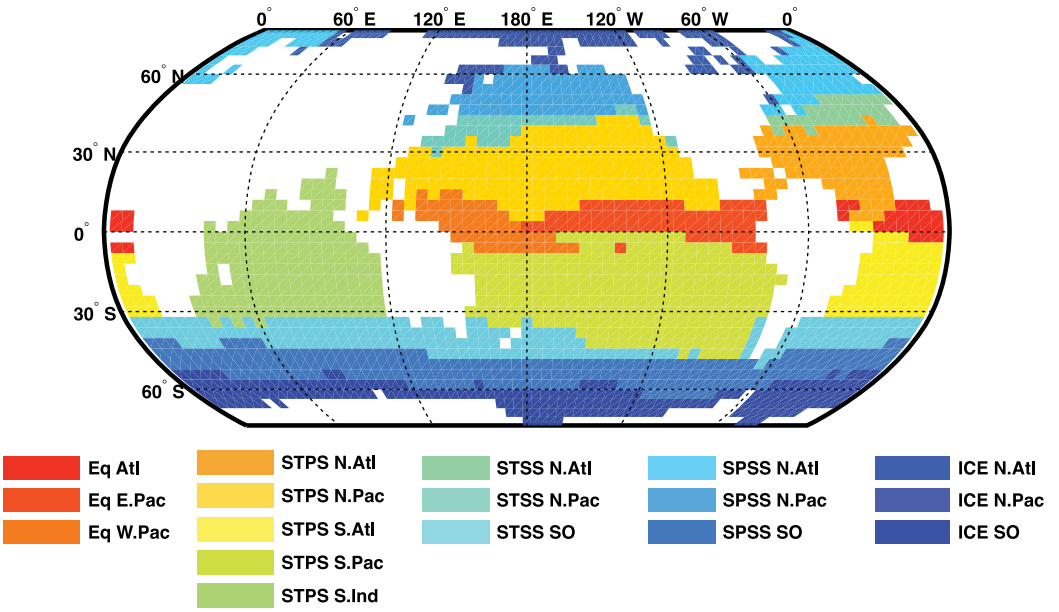

Figure 1. Global ocean divided into the biomes used in the present study, defined using the criteria outlined in *Fay and McKinley* [2014]. Sea surface temperature and ice data obtained from the Hadley Centre Sea Ice and Sea Surface Temperature data set, averaged over 1998–2010. Chlorophyll data obtained from the SeaWiFS sensor, also averaged over 1998–2010, and mixed layer depths obtained from the *Holte and Talley* [2009] climatology, which uses ARGO float data collected between 2002 and 2008. Eq refers to equatorial; STPS refers to subtropical permanently stratified; STSS refers to subtropical seasonally stratified; SPSS refers to subpolar seasonally stratified; ICE refers to ice covered.

models, which employ varied criteria to defined the MLD and do not all output the same set of variables, we use the fixed rather than variable biome definition.

Finally, we calculate the density of observations in each biome by dividing the number of observations from the *Takahashi et al*. [2009] database by the number of $4° \times 5°$ grid cells in each biome, to obtain the number of observations per grid cell.

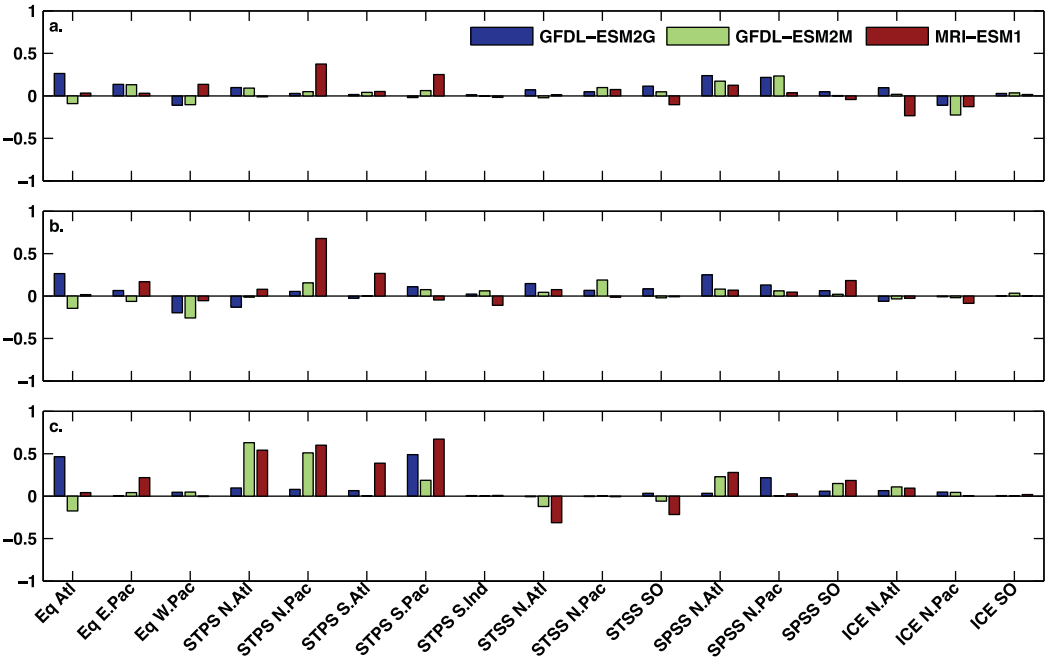

Figure 2. Model skill metrics created using biomes defined with observed SST and chlorophyll minus model skill metrics created using biomes defined with the SST and chlorophyll output from each model, for three representative models. Because each metric is on a [0, 1] scale, −1 and 1 represent the minimum and maximum of this difference.

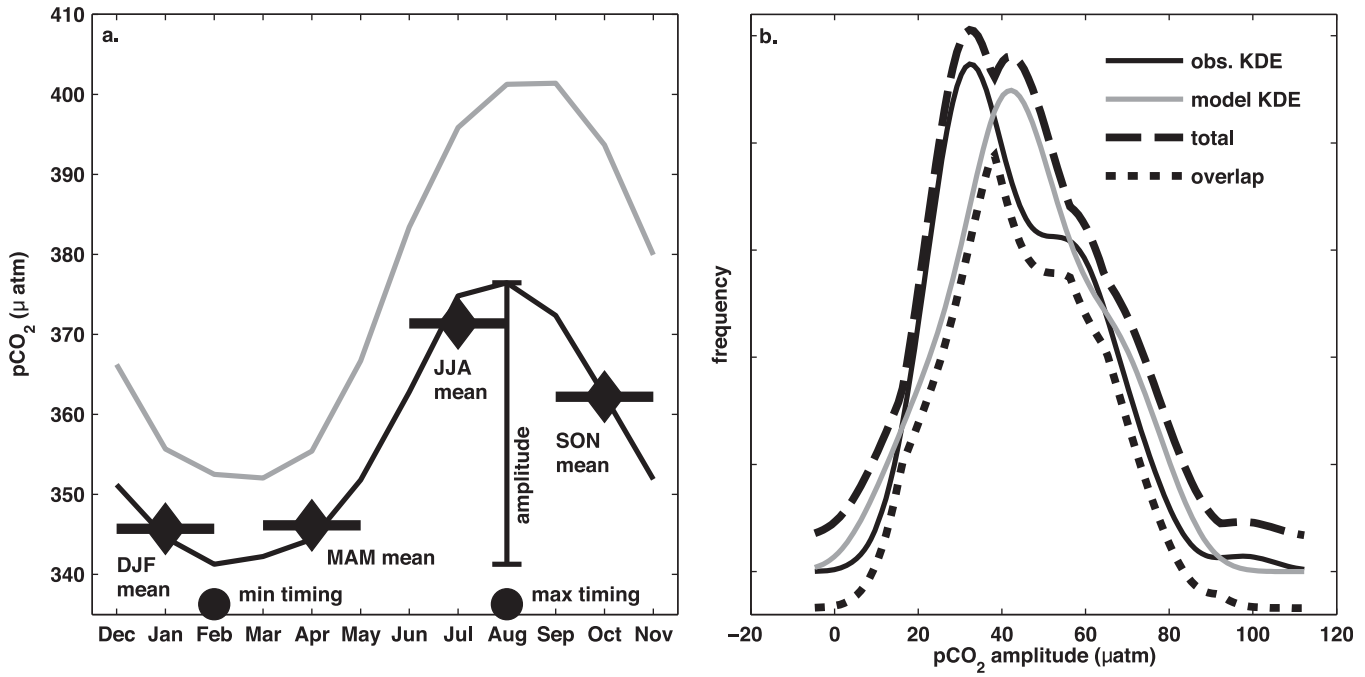

Figure 3. Schematic illustrating model skill metrics for the CESM model in the STSS North Atlantic. (a) Seasonal cycle of observed (black line) and CESM (gray line) pCO_2, averaged over the STSS-NA biome, with observed pCO_2 curve annotated to show the mean, amplitude, and timing metrics. (b) Kernel density estimates (KDEs) for the amplitude metric. The solid black and gray curves represent the observed and modeled KDEs, i.e., the distribution of pCO_2 seasonal amplitudes seen over all grid cells within the STSS-NA biome in the observational data and CESM model data. The black dashed curve represents the total area encompassed by both KDEs; including both the areas of overlap and the areas encompassed by only 1 KDE. The dotted curve represents the area of overlap between the KDEs only. The model skill metric is calculated by dividing the area under the overlap curve by the area under the total curve.

2.4. Model Metric

We create model metrics for seven variables describing the pCO_2^{SW} seasonal cycle: the mean pCO_2^{SW} over the spring (MAM), summer (JJA), fall (SON), and winter (DJF) seasons, the timing of the minimum pCO_2^{SW}, the timing of the maximum pCO_2^{SW}, and the amplitude of the yearly pCO_2^{SW} cycle (Figure 3a). For each variable and each biome, we calculate values for both modeled and observed pCO_2^{SW} at each $4^\circ \times 5^\circ$ grid cell within the biome. We then use all grid cells within each biome to calculate Kernel Density Estimates (KDEs) for the modeled and observed values. KDEs are discrete distributions, where each point is replaced by a distribution, or kernel, scaled to 1 and centered at the location of the original point [*Rosenblatt*, 1956; *Parzen*, 1962]. The KDEs smooth discrete distributions and incorporate the uncertainty inherent in each grid cell's value. We use the ksdensity Matlab function [*Bowman and Azzalini*, 1997] with a normal kernel to calculate the KDEs. The model metric for each category is then the degree of overlap between the model and the observational KDEs, found by dividing the area common to both KDE curves with the total area under the KDE curves. A value of 1 represents complete overlap, while a value of 0 represents no overlap (see schematic, Figure 3).

To further summarize model performance, we then calculate a seasonal mean metric as the average of the four seasonal metric values and a timing metric as the average of the minimum timing and maximum timing metric values at each biome. As a result, for each model and each biome, we determine three summary metrics for the pCO_2^{SW} seasonal cycle: seasonal mean, amplitude, and timing.

We also calculate model rankings in order to condense the information of all the metric skill scores for each model in each biome and to determine models that are generally performing the best across the entire domain. These rankings are calculated by taking the average skill scores of each metric across all ocean basins within each biome type (e.g., SPSS, STPS, ICE, etc.). This allows us to rank all 18 CMIP5 models for each metric and in each biome. Within these rankings, we group the models into separate "tiers" that distinguish between relative scores within the rankings. This tier structure results from the general observation that several models will perform comparatively, before a substantial jump down to the next grouping of models. Models are grouped into the same tier if their skill score is within 5% of the next highest ranked model.

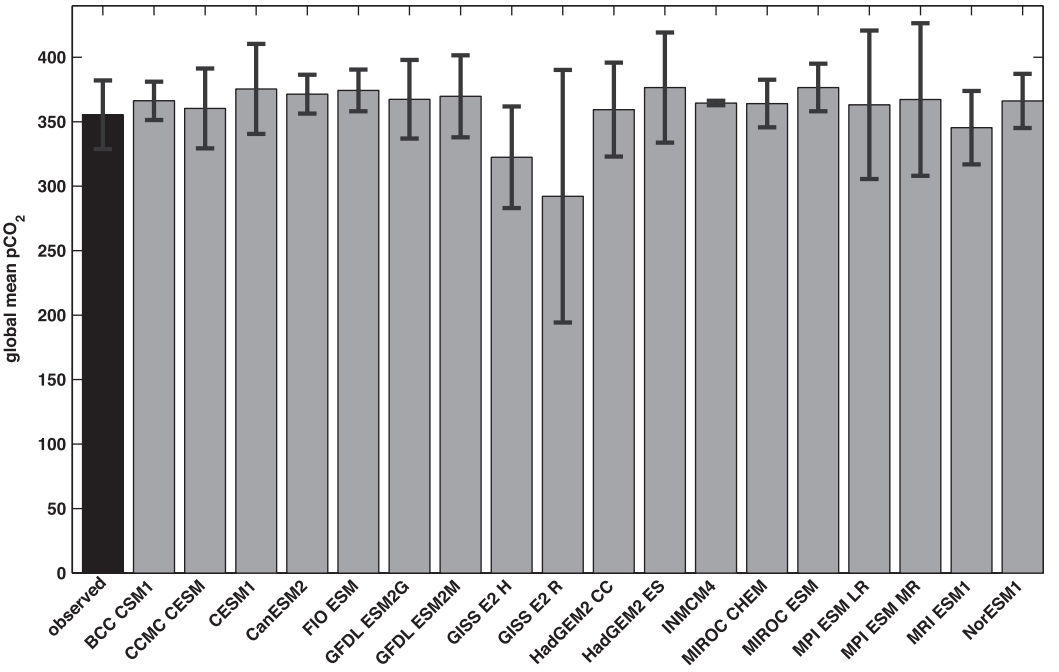

Figure 4. Global mean annual average pCO_2^{SW} in the observed *Takahashi et al.* [2009] climatology and the CMIP5 models. Errorbars denote the spatial variability across the biomes as the 1 sigma standard deviation.

2.5. CESM Large Ensemble

In an effort to distinguish between model metric variability due to mechanistic differences within the CMIP5 spread and variability due to internal variability, we utilize 24 ensemble members from the Community Earth System Model Large Ensemble (CESM-LE) project [*Kay et al.*, 2015]. Although this version of the CESM is similar to the version submitted to CMIP5, the CESM-LE uses the updated atmosphere model CAM5, as opposed to CAM4 in CMIP5. Output from the CESM-LE is available via the Earth System Grid (https://www.earthsystemgrid.org/home.htm). Each ensemble member consists of the same model with identical external forcing, but is generated with a slight perturbation ($\sim 10^{-14}$ K) to the initial air temperature field. This sets each ensemble member on a unique climate trajectory. Thus, differences in the evolution of the Earth system in each ensemble member are taken to be representative of internal variability within the modeled climate system.

For each ensemble member, we create a monthly pCO_2^{SW} climatology using monthly pCO_2^{SW} data from the 11 year timeframe between 1995 and 2005. Each climatology is then regridded to the $4^\circ \times 5^\circ$ spatial resolution using the same methodology as the CMIP5 model climatologies. Next, model metrics are calculated using the same methodology as for the CMIP5 models. Due to the difference in the number of simulations analyzed between the CMIP5 and CESM-LE projects (18 versus 24, respectively), we utilize a Monte-Carlo approach to randomly sample 18 of the 24 CESM-LE members. From these 18 randomly sampled members, the standard deviation across the ensemble is computed for each metric and biome. The Monte-Carlo simulation was run 1000 times, at which point the maximum change in the standard deviation for any given metric or biome was $<0.1\%$. Thus, we consider the magnitude of the variability in model metric skill score computed across the CESM-LE to be representative of the differences between CMIP5 model metrics that may be attributable to internal variability.

3. Results

3.1. Overview of Metrics and Model Performance

Figure 4 illustrates how the CMIP5 models compare to the observations when examining annual, globally averaged pCO_2^{SW}. Most CMIP5 models fall within 50 μatm of the observed average value, with the majority slightly greater than the observed. The variability across the biomes is also similar in many of the CMIP5

models, though a few models contain substantially greater variability (e.g., GISS-E2-R and the MPI models), while the INMCM4 model contains substantially lower variability. Thus, many of the CMIP5 models simulate annual, globally averaged pCO_2^{SW} reasonably well, with comparable biome-scale spatial variability.

The results of all model metric scores in all ocean biomes are summarized in Figure 5. In most biomes and for most metrics, model skill scores tend to vary by up to 0.5, with somewhat lower variance in the timing metric. The variability across both biomes and metrics makes it difficult to determine the relative performance of models based on the raw metric scores alone; therefore, we calculate model rankings and divide models into separate tiers based on average performance (Table 2). The GFDL models contain the greatest number of average skill scores in the top tiers, with particularly strong average scores in the SPSS biomes. The CanESM2 also falls within the top tiers for many of the biomes. The CESM1 generally falls in the top tiers for the amplitude and timing metrics, but does not perform as well for the mean metric. Some models such as BCC CSM1 top the rankings for a specific metric within a specific region (e.g., mean STPS and ICE), but do not score as well in other categories.

We conducted an additional assessment to test for evidence of model tuning, by examining whether models that performed significantly better in any biome than the mean of all models ($>$1 standard deviation) then performed worse than the mean of metrics for all other biomes. We only found one instance in which a model with unusually good performance in one region performed worse than the mean in more than half of the other regions (HadGEM2-CC for the timing metric in the North Pacific ICE biome) so do not see widespread evidence of model tuning.

3.2. Comparison Across Biomes

Figure 6 shows model metric scores for each biome, averaged across all 18 CMIP5 models used. The mean and amplitude metric scores are generally greater in biomes located within the subtropical and midlatitude regions (0.3–0.5), while metric scores in equatorial and polar regions are generally lower (0.2–0.4 and 0.2–0.3, respectively). This pattern does not hold in the timing metric, as the greatest scores are in the equatorial and subtropical biomes. Taken as a collective group, the CMIP5 models perform the best at the timing metric, with metric skill scores that are generally twice as greater than the amplitude and mean metric skill scores. The amplitude and mean metric skill scores are comparable for the collective CMIP5 models, with the amplitude metric scoring slightly better overall.

3.3. Comparison with Density of Observations

We compute correlation coefficients between model skill score and observational data density within each biome for the 18 CMIP5 models to determine the impact that the number of observational data points has on model skill score (Figure 7). We find that model performance is generally weakly correlated with the density of observations in each biome. Mean metric skill scores display a weak negative correlation with observational data density (i.e., lower model performance in regions with more observations). Correlation coefficients between -0.25 and -0.50 occur most frequently, and all correlation coefficients fall between -0.75 and 0 for the mean metric. Conversely, the timing and amplitude metrics display weak positive correlations with data density (i.e., higher model performance with more observations). Correlation coefficients fall between -0.50 and 0.75 with the greatest frequency between 0 and 0.25 for the amplitude metric. For the timing metric, correlation coefficients fall between -0.50 and 0.50 with the greatest frequency between 0 and 0.25.

3.4. Assessing the Contribution of Interannual Variability

The intermodel variability across the CMIP5 models is compared to the variability in a large ensemble of a single model (CESM-LE) in Figure 8. Variability is described as the standard deviation for the two model ensembles. Variability across CMIP5 is considerably greater than variability within the CESM-LE. For the mean metric, CESM-LE variability is generally less than 10% of the CMIP5 variability, except in the Equatorial West Pacific biome where CESM-LE variability is greater than 25% of the CMIP5 variability. CESM-LE variability is somewhat greater in the amplitude metric and is comparable to greater than 10% of the CMIP5 variability in roughly half of the biomes. CESM-LE variability is greatest in the timing metric, while CMIP5 variability is smallest. Therefore, for selected regions, notably the STPS South Atlantic and STSS North Atlantic, CESM-LE variability in the timing metric is greater than 20% of the CMIP5 variability. A few biomes, such as the Equatorial East and West Pacific, have relatively greater CESM-LE variability for all three model

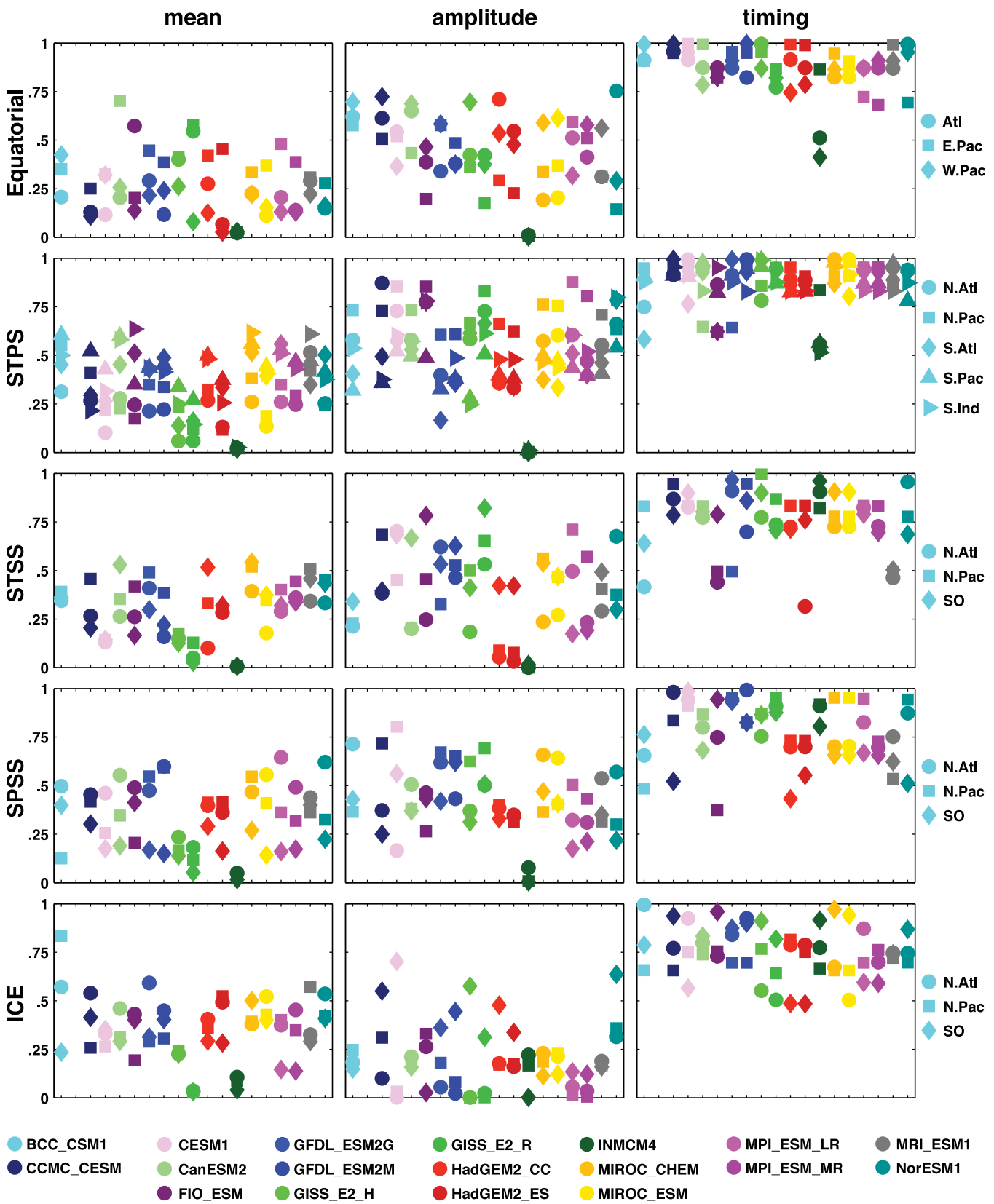

Figure 5. Summary of metrics scores for all biomes and models. Each plot represents the metric scores of all models in a biome environment and for each type of metric. The marker colors and *x* axis in each plot represent the different models, the *y* axis represents the metric score (on a scale of 0–1 for all plots), and the different markers represent the ocean basins within each biome type.

Table 2. Ranking of Top 5 Models for Each Metric and Biome Type[a]

EQ	STPS	STSS	SPSS	ICE
Mean				
1. GISS-E2-R	1. BCC-CSM1	1. MIROC-CHEM	1. GFDL-ESM2M	1. BCC-CSM1
1. CanESM2	1. MIROC-CHEM	2. GFDL-ESM2G	1. MIROC-CHEM	2. MIROC-ESM
2. GISS-E2-H	2. MPI-ESM-LR	2. CanESM2	2. GFDL-ESM2G	2. HadGEM2-ES
3. BCC-CSM1	2. CanESM2	2. MPI-ESM-MR	2. CCMC-CESM	2. MIROC-CHEM
3. GFDL-ESM2G	2. HadGEM2-CC	2. BCC-CSM1	2. MPI-ESM-LR	2. CCMC-CESM
Amplitude				
1. BCC-CSM1	1. FIO-ESM	1. GISS-E2-R	1. GFDL-ESM2G	1. NorESM1
1. CCMC-CESM	2. NorESM1	2. CESM1	1. GFDL-ESM2M	2. CCMC-CESM
1. CanESM2	2. GISS-ES-R	3. GFDL-ESM2M	1. GISS-E2-R	3. HadGEM2-CC
2. HadGEM2-CC	2. CESM1	4. FIO-ESM	2. CESM1	4. CESM1
2. GFDL-ESM2G	3. MPI-ESM-LR	4. GFDL-ESM2G	2. BCC-CSM1	5. HadGEM2-ES
Timing				
1. CCMC-CESM	1. CCMC-CESM	1. INMCM4	1. CESM1	1. GFDL-ESM2M
1. CESM1	1. MIROC-CHEM	1. GISS-E2-H	1. GFDL-ESM2G	1. FIO-ESM
1. GISS-E2-H	1. GFDL-ESM2M	1. CCMC-CESM	1. GISS-E2-R	1. BCC-CSM1
1. BCC-CSM1	1. MIROC-ESM	1. CESM1	1. GFDL-ESM2M	1. GFDL ESM2G
1. GFDL-ESM2M	1. GISS-E2-R	1. GFDL-ESM2M	1. INMCM4	1. CanESM2

[a]For each type of biome, metrics are multiplied together across ocean basins and northern/southern latitudes. Models that score within 5% of each other are classed in the same tier, as indicated by the numbers.

metrics. Other biomes, such as the STPS South Atlantic, have moderately high CESM-LE variability for one metric (timing), but relatively low variability for the other metrics.

4. Discussion

In this study, we assessed the ability of CMIP5 models to simulate the seasonal cycle of pCO_2^{SW} in different regions based on a comparison of the seasonal mean, amplitude, and timing with an observational climatology. We expand on previous work [e.g., *Anav et al*., 2013] by focusing on the seasonality of a specific variable, the use of biogeographically constrained ocean biomes, comparisons to observational density, and a quantitative analysis of the impact of internal variability on model skill scores.

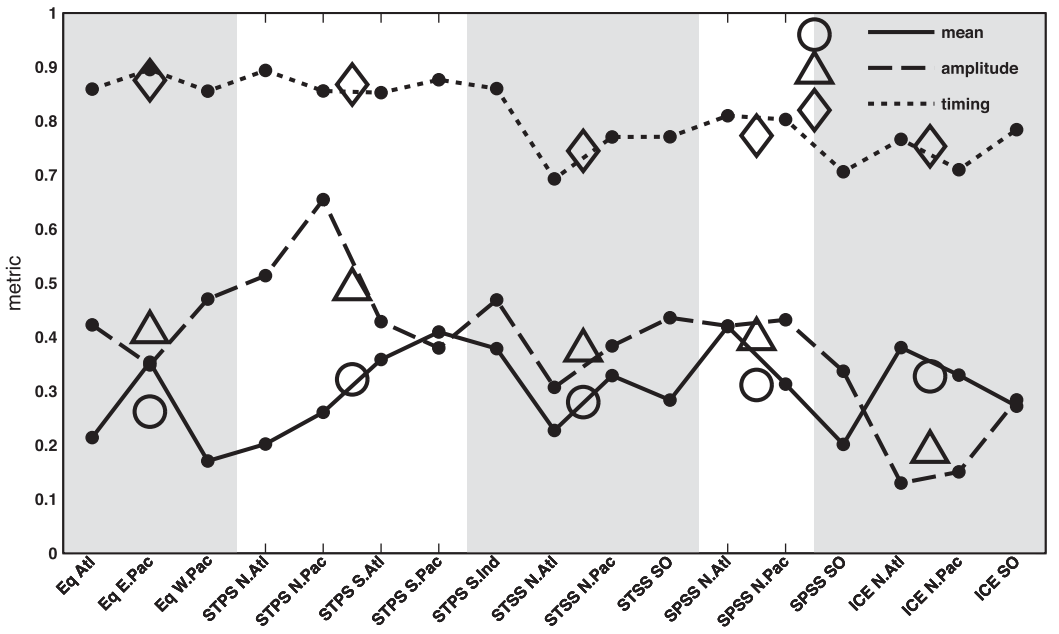

Figure 6. Comparison of model skill metrics, averaged over the 18 models, for each biome. The lines and filled circles represent the average mean, amplitude, and timing metric for each biome. The open symbols represent the mean, amplitude, and timing metrics further averaged over the type of biome (Eq, STPS, STSS, SPSS, and ICE). The shading separates the different types of biomes.

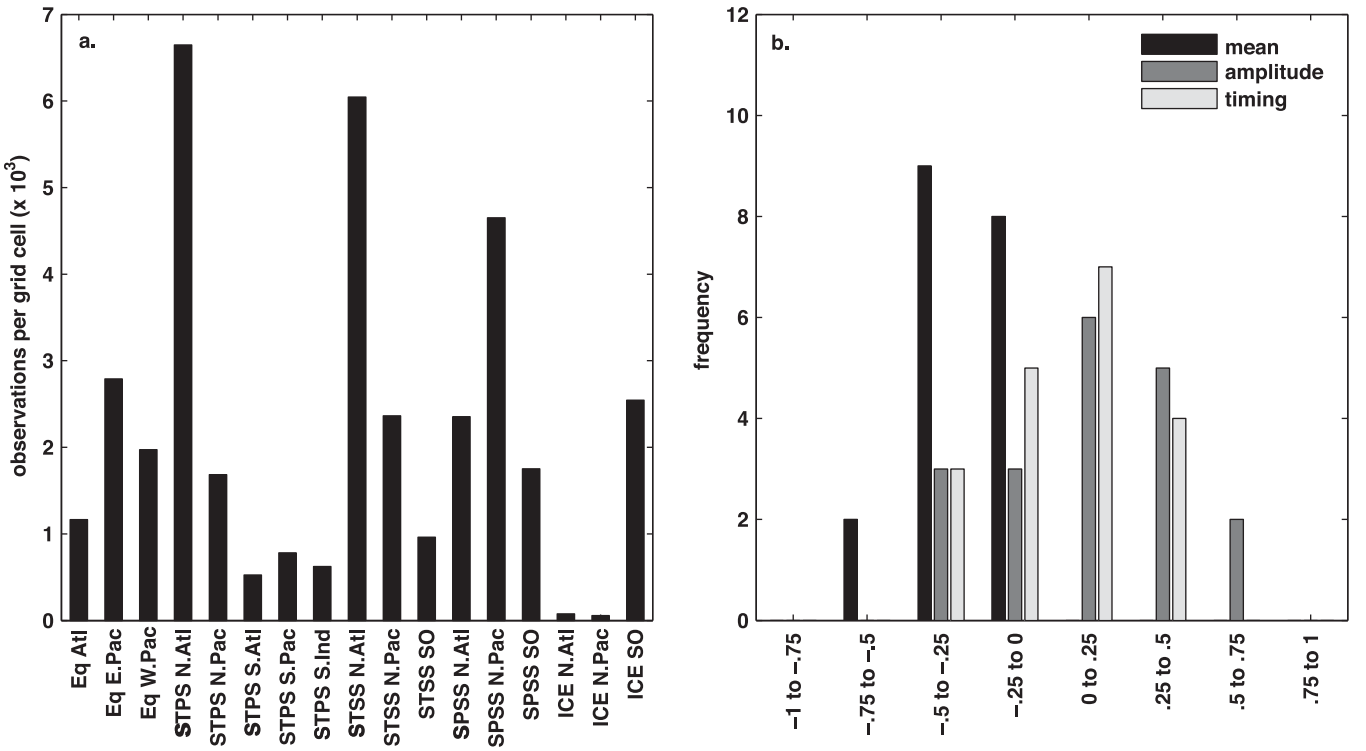

Figure 7. (a) Average number of observations per grid cell for each biome. (b) Histogram of the correlation between model metrics for each biome versus the density of observations in each biome, for all 18 models. The histogram for each metric is shown separately.

The model skill scores and rankings presented show that the models are most accurate when simulating the timing of the seasonal cycle, and least accurate when simulating the overall mean. However, because models can achieve a timing skill score of 1 if they correctly reproduce the month of the pCO_2^{SW} minimum

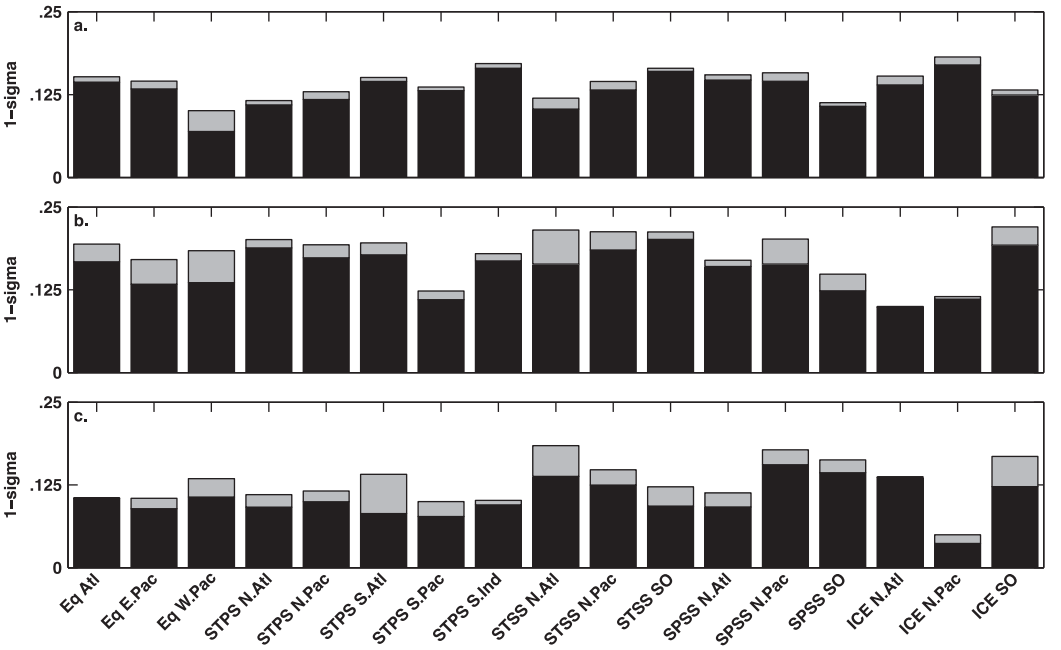

Figure 8. Comparison of CMIP5 model variability versus CESM ensemble member variability for each metric (mean, amplitude, and timing) and biome. The total height of each bar represents the standard deviation (1 sigma) among the 18 CMIP5 models for each metric and biome. The gray shading at the top of each bar represents the average standard deviation among 18 CESM large ensemble members, calculated using a Monte-Carlo simulation of successive randomly selected groups of 18 ensemble members.

and maximum in each grid cell of a biome, whereas models can only achieve an amplitude or mean metric score of 1 if they reproduce the exact value of mean/amplitude pCO_2^{SW} in each region, the higher timing skill scores can be expected. Although the models perform the worst on the mean skill score, many of the models produce annual, globally averaged pCO_2^{SW} values that are comparable to the observations. However, it is possible for a model to output a comparable annual pCO_2^{SW} value, but simulate a seasonal cycle that differs greatly from the observations, and thus perform poorly in our skill score metric. This distinction has potential implications for CO_2 flux and ocean carbon uptake, since the flux is also dependent on seasonally varying wind speed.

The GFDL models score well across many of the metrics and regions, consistent with the high skill scores of the GFDL models in simulating ocean CO_2 flux in *Anav et al*. [2013]. Yet there are still biomes such as the STPS and ICE where the GFDL models do not rank as high. The MIROC, BCC CSM1, CanESM2, GISS, CESM1, and CCMC CESM also perform well for particular metrics or biomes. These results make it difficult to denote a particular model as the most accurate, considering that this grouping of models all rank highly. Additionally, we do not find evidence that models have been tuned to better perform in a given biome or for a given aspect of the seasonal cycle.

A comparison of average model skill scores across biomes reveals that models generally perform best in midlatitude regions (STPS, STSS, SPSS) and worst in the equatorial and ICE biomes. The poor performance in the equatorial region is at least partly due to the exclusion of El Niño years in the observational climatology, while similar El Niño patterns are retained in the models (if they occurred during the 1995–2005 timeframe). Furthermore, within each type of biome, there can be a wide range in the density of observations used to compute the pCO_2^{SW} data set (e.g., the STPS North Atlantic has a higher density of observations than the STPS South Atlantic by a factor of >10). It might be expected that in regions with more observations and better-constrained estimates of the pCO_2^{SW} seasonal cycle, models would perform better, due to more sufficient knowledge of the dominant physical and biological processes for models to incorporate. This appears to be the case for the timing and amplitude metrics, which display weakly positive correlations to the density of observations. However, the mean metric is weakly negatively correlated with observational density, denoting slightly worse model performance in areas with more measurements per grid cell. Thus, for the seasonal mean metric, the greater density of observations may provide sufficient process information with which to invalidate the models. However, the seasonal mean, timing, and amplitude metrics are generally very weakly correlated with model density ($r < 0.25$ for over 50% of models); therefore, these proposed explanations may not play a large role in explaining differences in model skill between biomes.

Taken together, these findings imply that model performance is more sensitive to the physics and biology controlling pCO_2^{SW} in a given region of the ocean, rather than the degree to which the pCO_2^{SW} seasonal cycle has been studied and constrained. Because our skill scores do not assess a model's ability to simulate important physical or biological ocean variables, such as sea surface temperature, chlorophyll, or dissolved inorganic carbon concentration, it is probable that model deficiencies in simulating pCO_2 in a given region can be traced to deficiencies in one or more of these variables due to missing or poorly constrained processes in the models. For example, while model skill is generally high in the SPSS biomes, models perform poorly in the Southern Ocean SPSS biome compared with the North Atlantic and North Pacific SPSS biomes, and compared with average model performance globally for all three metrics. Inaccuracies in modeled physical processes (e.g., upwelling and outgassing of CO_2) due to the location and strength of the westerly winds in the Southern Ocean, a known issue in CMIP5 models [*Bracegirdle et al*., 2013], provide a potential explanation for this relatively poor performance. Conversely, strong model scores may be connected to simulating a particular field accurately. While this detailed analysis is left to future work, we note the potential for the model skill metrics developed here to highlight specific processes relevant to the global carbon cycle that should be further developed and improved in the CMIP5 models.

The lack of a relationship between observational density and model skill score further implies that our calculated model skill scores are relatively insensitive to our choice of the observational data set selected. However, we note that other pCO_2 data sets have become available that offer greater spatial and temporal resolution than the *Takahashi et al*. [2009] climatology that is used here and is widely used in the research community [e.g., *Anav et al*., 2013; *Fay and McKinley*, 2013; *Long et al*., 2013]. Version 2 of the SOCAT database [*Bakker et al*., 2014] contains approximately 10.1 million measurements of fCO_2 compared to the approximately 3 million measurements of pCO_2 in the *Takahashi et al*. [2009] data set. An additional

climatology using a two-step neural network technique with the SOCAT database was released at 1° × 1° resolution [*Landschützer et al.*, 2014a, 2014b]. If these additions and modifications result in significant differences in the value of pCO_2^{SW} regionally or seasonally, we might expect our model skill scores to change when calculated with a different observational data set.

Model internal variability does not appear to significantly affect our assessment of model skill. This is most likely because we compared modeled and observed pCO_2^{SW} climatologies, created from a 10-plus year timeframe. We therefore conclude that assessments of the seasonal cycles of ESM variables are likely consistent across model ensembles and not substantially impacted by internal variability, provided that a multiyear climatology is used to examine the seasonal cycle. However, we note that our quantitative analysis of internal variability is based entirely on the CESM-LE, since this is one of the few CMIP5 models that have completed a large ensemble experiment. It is unknown how the internal variability within the CESM compares to other CMIP5 models. For example, ENSO cycles in CCSM4 (the precursor of the CESM) are approximately 30% greater in magnitude compared to observed ENSO cycles [*Deser et al.*, 2012b]. If other CMIP5 models have an ENSO cycle closer to observed, then they may contain a lower magnitude of internal variability compared to the CESM.

Our model rankings are not presented as definitive support of a particular model's skill or predictive power in simulating pCO_2^{SW}. Rather, our goal is to present a novel method for calculating model skill scores that can provide vital information for researchers deciding which model they may wish to select for a particular research study. Our results suggest that certain models may be better suited for different research questions, even when exclusively considering a single variable, and can inform the choice of CMIP5 model used for future research. Additionally, this method can be exported to the assessment of the seasonal cycle of other CMIP5 ESM variables with a global observational database to shed further light on global model performance and use for predicting large-scale changes in the global carbon cycle.

Acknowledgments

This work was initiated as part of the National Center for Atmospheric Research Advanced Study Program (NCAR ASP) Summer Colloquium 2013, "Carbon-Climate Connections in the Earth System." All of the data used in this analysis are available online. The CMIP5 model data are available via the Program for Climate Model Diagnosis and Intercomparison (PCMDI) website (http://cmip-pcmdi.llnl.gov/cmip5/). The CESM-LE model data are available from the Earth System Grid website (https://www.earthsystemgrid.org/home.htm). The *Fay and McKinley* [2014] biome definitions are available from the PANGAEA Data Publisher for Earth and Environmental Science website (http://doi.pangaea.de/10.1594/PANGAEA.828650). We thank two anonymous reviewers and the Editor for their constructive comments that helped improve this manuscript. We thank Juan Muglia for assisting during the ASP colloquium. We thank the modeling centers, the PCMDI, and the WCRP for making the model output data readily available. Special thanks to the NCAR ASP program for funding and to Matt Long (NCAR) and Galen McKinley (University of Wisconsin-Madison) for their constructive comments and encouragement.

References

Anav, A., P. Friedlingstein, M. Kidston, L. Bopp, P. Ciais, P. Cox, C. Jones, M. Jung, R. Myneni, and Z. Zhu (2013), Evaluating the land and ocean components of the global carbon cycle in the CMIP5 Earth System Models, *J. Clim.*, *26*, 6801–6843, doi:10.1175/JCLI-D-12-00417.1.

Bakker, D. C. E., et al. (2014), An update to the Surface Ocean CO_2 Atlas (SOCAT version 2), *Earth Syst. Sci. Data*, *6*, 69–90, doi:10.5194/essd-6-69-2014.

Barnes, S. L. (1994), Applications of the Barnes objective analysis scheme. Part I: Effects of undersampling, wave position, and station randomness, *J. Atmos. Oceanic Technol.*, *11*, 1433–1448.

Bowman, A. W., and A. Azzalini (1997), *Applied Smoothing Techniques for Data Analysis*, Oxford Univ. Press, N. Y.

Bracegirdle, T. J., E. Shuckburgh, J.-B. Sallee, Z. Wang, A. J. S. Meijers, N. Bruneau, T. Phillips, and L. J. Wilcox (2013), Assessment of surface winds over the Atlantic, Indian, and Pacific Ocean sectors of the Southern Ocean in CMIP5 models: Historical bias, forcing response, and state dependence, *J. Geophys. Res. Atmos.*, *118*, 547–562, doi:10.1002/jgrd.50153.

Ciais, P., et al. (2013), *Carbon and other biogeochemical cycles, in Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by T. F. Stocker et al., chap. 6, 1535 pp., Cambridge Univ. Press, Cambridge, U. K.

Deser, C., A. Phillips, V. Bourdette, and H. Teng (2012a), Uncertainty in climate change projections: The role of internal variability, *Clim. Dyn.*, *38*, 527–546, doi:10.1007/s00382-010-0977-x.

Deser, C., A. Phillips, R. A. Tomas, Y. M. Okumura, M. A. Alexander, A. Capotondi, J. D. Scott, Y.-O. Kwon, and M. Ohba (2012b), ENSO and Pacific decadal variability in the community climate system model version 4, *J. Clim.*, *25*, 2622–2651, doi:10.1175/JCLI-D-11-00301.1.

Fay, A. R., and G. A. McKinley (2013), Global trends in surface ocean pCO2 from in situ data, *Global Biogeochem. Cycles*, *27*, 541–557, doi:10.1002/gbc.20051.

Fay, A. R., and G. A. McKinley (2014), Global ocean biomes: Mean and temporal variability, *Earth Syst. Sci. Data*, *6*, 273–284, doi:10.5194/essd-6-273-2014.

Friedlingstein, P., et al. (2006), Climate-carbon cycle feedback analysis: Results from the C^4MIP Model intercomparison, *J. Clim.*, *19*, 3337–3353.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux (2008), Performance metrics for climate models, *J. Geophys. Res.*, *113*, D06104, doi:10.1029/2007JD008972.

Hauck, J., and C. Völker (2015), Rising atmospheric CO_2 leads to large impact of biology on Southern Ocean CO_2 uptake via changes of the Revelle factor, *Geophys. Res. Lett.*, *42*, 1459–1464, doi:10.1002/2015GL063070.

Holte, J., and L. Talley (2009), A new algorithm for finding mixed layer depths with applications to argo data and subantarctic mode water formation*, *J. Atmos. Oceanic Technol.*, *26*, 1920–1939, doi:10.1175/2009JTECHO543.1.

Kay, J. E., et al. (2015), The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability, *Bull. Am. Meteorol. Soc.*, doi:10.1175/BAMS-D-13-00255.1.

Khatiwala, S., F. Primeau, and T. Hall (2009), Reconstruction of the history of anthropogenic CO_2 concentrations in the ocean, *Nature*, *462*(7271), 346–349.

Landschützer, P., N. Gruber, D. C. E. Bakker, and U. Schuster (2014a), An observation-based global monthly gridded sea surface pCO_2 product from 1998 through 2011 and its monthly climatology, Carbon Dioxide Inf. Anal. Cent., Oak Ridge Natl. Lab., U.S. Dep. of Energy, Oak

Ridge, Tenn., doi:10.3334/CDIAC/OTG.SPCO2_1998_2011_ETH_ SOM-FFN. [Available at http://cdiac.ornl.gov/ftp/oceans/spco2_1998_2011_ETH_SOM-FFN.]

Landschchützer, P., N. Gruber, D. C. E. Bakker, and U. Schuster (2014b), Recent variability of the global ocean carbon sink, *Global Biogeochem. Cycles*, *28*, 927–949, doi: 10.1002/2014GB004853.

Lin, J. L. (2007), Interdecadal variability of ENSO in 21 IPCC AR4 coupled GCMs, *Geophys. Res. Lett.*, *34*, L12702, doi: 10.1029/2006GL028937.

Lindsay, K., G. Bonan, S. Doney, F. Hoffman, D. Lawrence, M. Long, N. Mahowald, J. Moore, J. Randerson, and P. Thornton (2014), Preindustrial control and 20th century carbon cycle experiments with the earth system model CESM1(BGC), *J. Clim.*, *27*, 8981–9005, doi:10.1175/JCLI-D-12-00565.1.

Long, M. C., K. Lindsay, S. Peacock, J. K. Moore, and S. C. Doney (2013), Twentieth-century oceanic carbon uptake and storage in CESM1(BGC), *J. Clim.*, *26*, 6775–6800, doi:10.1175/JCLI-D-12-00184.1.

Parzen, E. (1962), On estimation of a probability density function and mode, *Ann. Math. Stat.*, *33*(3), 1065–1076, doi:10.1214/aoms/1177704472.

Radić, V., and G. K. C. Clarke (2011), Evaluation of IPCC models' performance in simulating late-twentieth-century climatologies and weather patterns over North America, *J. Clim.*, *24*, 5257–5274, doi:10.1175/JCLI-D-11-00011.1.

Rodgers, K. B., J. L. Sarmiento, O. Aumont, C. Crevoisier, C. de Boyer Montégut, and N. Metzl (2008), A wintertime uptake window for anthropogenic CO_2 in the North Pacific, *Global Biogeochem. Cycles*, *22*, GB2020, doi:10.1029/2006GB002920.

Rosenblatt, M. (1956), Remarks on some nonparametric estimates of a density function, *Ann. Math. Stat.*, *27*(3), 832–837, doi:10.1214/aoms/1177728190.

Scherrer, S. C. (2011), Present-day interannual variability of surface climate in CMIP3 models and its relation to future warming, *Int. J. Climatol.*, *31*, 1518–1529, doi:10.1002/joc.2170.

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

Takahashi, T., et al. (2009), Climatological mean and decadal change in surface ocean pCO_2, and net sea–air CO_2 flux over the global oceans, *Deep Sea Res., Part II*, *56*(8), 554–577, doi:10.1016/j.dsr2.2008.12.009.

Taylor, K. E., R. J. Stouffer, and G. Meehl (2011), An overview of CMIP5 and the experiment design, *Bull. Am. Meteorol. Soc.*, *93*, 485–498.