Evaluating the ocean biogeochemical components of earth system models using atmospheric potential oxygen (APO) and ocean color data

C.D. Nevison[1], M. Manizza[2], R.F. Keeling[2], M. Kahru[2], L. Bopp[3], J. Dunne[4], J. Tiputra[5], T. Ilyina[6], B.G. Mitchell[2]

[1] University of Colorado, Boulder, Institute for Arctic and Alpine Research, Boulder, Colorado

[2] Scripps Institution of Oceanography, La Jolla, California.

[3] IPSL/LSCE, UMR8212, CNRS-CEA-UVSQ, Gif sur Yvette, France

[4] National Oceanic and Atmospheric Administration/Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey

[5] Uni Climate, Uni Research and Bjerknes Centre for Climate Research, Bergen, Norway.

[6] Max Planck Institute for Meteorology, Hamburg, Germany.

Abstract

The observed seasonal cycles in atmospheric potential oxygen (APO) at a range of mid to high latitude surface monitoring sites are compared to those inferred from the output of 6 Earth System Models participating in the fifth phase of the Coupled Model Intercomparison Project (CMIP5). The simulated air-sea O_2 fluxes are translated into APO seasonal cycles using a matrix method that takes into account atmospheric transport model (ATM) uncertainty among 13 different ATMs. Three of the ocean biogeochemistry models tested are able to reproduce the observed APO cycles at most sites, to within the large TransCom3-era ATM uncertainty used here, while the other three generally are not. Net Primary Production (NPP) and net community production (NCP), as estimated from satellite ocean color data, provide additional constraints, albeit more with respect to the seasonal phasing of ocean model productivity than overall magnitude. The present analysis suggests that, of the tested ocean biogeochemistry models, CESM and GFDL ESM2M are best able to capture the observed APO seasonal cycle at both Northern and Southern Hemisphere sites. Uncertainties in most models can be attributed to the underestimation of NPP, deep ventilation or both in the northern oceans.

Introduction

Ocean physical and biogeochemical processes have profound influences on Earth's climate. Phytoplankton in the sunlit part of the ocean convert carbon from inorganic to organic form via photosynthesis, thereby establishing the base of the ocean food chain. Primary production and subsequent export of organic carbon from the mixed layer (export production) and remineralization at depth are key components of the so called "biological pump," which regulates the partition of carbon between the ocean and atmosphere [*Gruber and Sarmiento*, 2002; *Boyd & Doney*, 2003].

Net community production (NCP) and the related process of export production (EP) are also important for understanding the distribution of dissolved O_2 within the ocean and the flux of O_2 (F_{O2}) at the air-sea interface. NCP is defined here as the net amount of organic carbon fixed through photosynthesis over the depth of the mixed layer after accounting for grazing and both autotrophic and heterotrophic respiration. NCP is closely linked to F_{O2}, since each mole of photosynthetically-fixed carbon that persists beyond the time scale of air-sea exchange (2-3 weeks) leaves a stoichiometric amount of O_2 available for release to the atmosphere. This release of O_2 to the atmosphere in association with NCP occurs mainly in the spring and summer at extratropical latitudes [*Keeling et al.*, 1993]. EP more or less balances NCP when averaged over a full year or if the upper ocean is in a long-term steady state and advective fluxes are zero [*Laws et al.*, 2000]. The exported carbon subsequently is respired in the subsurface ocean, leading to O_2 depletion at depth. O_2 is replenished by absorption from the atmosphere when the deep waters mix back to the surface in fall and winter. Deep ventilation and NCP thus are distinct processes that are largely separate in time and space but are both closely linked to the biological pump critical that draws carbon out of surface waters and is critical for ocean uptake of atmospheric CO_2.

To explore the impacts of future climate change on Earth's climate and ecosystems, the Coupled Model Intercomparison Project phase 5 (CMIP5) relies on 3-dimensional numerical Earth System Models (ESMs), which incorporate descriptions of biogeochemical impacts of land and marine biota. Projections of future atmospheric CO_2 levels and associated climate warming in CMIP5 depend not only on fossil fuel use projections but also on assumptions about uptake and storage of carbon by the land and ocean. The oceans have absorbed approximately one third of the anthropogenic carbon released to the atmosphere since the beginning of the industrial era [*Khatiwala et al.*, 2009], but this fractional rate of uptake is unlikely to continue in the future as the buffering capacity of surface waters declines and the export of carbon from the surface to the deep ocean fails to keep pace with anthropogenic fossil fuel combustion [*Arora et al.*, 2013]. Changes in ventilation of abyssal deepwater are an additional possible consequence of future climate forcing that current models may or may not be able to predict accurately [*Sigman et al.*, 2010].

Recent studies have tested the present-day skill of the ocean components of ESMs and some have also examined future projections [*Schneider et al.*, 2008; *Steinacher et al.*, 2010, *Bopp et al.* 2013; *Anav et al*, 2013]. These evaluations have compared model output to both hydrographic measurements and remotely sensed ocean color products, most commonly net primary production (NPP). The models predict spatial-annual patterns in NPP that reproduce some of the main features seen in satellite data, but differ over a factor of 2 in NPP magnitude. Some evaluations have examined seasonal variability and have found that ocean models tend to underestimate observed NPP at high latitudes (poleward of 44º) in the Northern Hemisphere and overestimate it in the Southern Hemisphere. The models also fail to capture the timing of the observed high latitude peak in NPP in both hemispheres, with predictions that are often 1-2 months earlier than observations [*Anav et al.*, 2013; *Henson et al.*, 2013]. However, ocean color-derived NPP values are uncertain, especially in the Southern Ocean, reducing confidence in the "observed" constraints.

Many biogeochemical processes that are expected to occur in the future, such as responses to warming and stratification, are also highly relevant on seasonal time scales [Keeling et al., 2010;

Anav et al., 2013]. Thus, challenging models against known seasonal variations can aid in the development of credible predictions of future changes. Here, we evaluate 6 earth system models used in CMIP5 against two cross-cutting metrics, which test the models' ability to account for changes in ocean biogeochemistry on seasonal time frames. This work is intended primarily as a demonstration of method using an available subset of the CMIP5 ESMs rather than as a comprehensive evaluation of all the CMIP5 models. The first metric is based on satellite-derived estimates of ocean color, focusing on NPP and NCP. The second metric is based on the seasonal cycles in atmospheric potential oxygen (APO), an atmospheric tracer that varies seasonally mainly due to air-sea exchanges of O_2 [*Stephens et al.*, 1998; *Manning and Keeling*, 2006].

NCP is the ocean color-derived flux most relevant to the biological pump, but cannot be directly observed by remote sensing. It is derived by a combination of remote measurements and poorly constrained models, which inherently increases its uncertainty [*Schneider et al.*, 2008; *Nevison et al.*, 2012a]. The quantity actually observed from space is spectral top of the atmosphere radiance, which is used to estimate chlorophyll (or another proxy of phytoplankton biomass); chlorophyll and other variables such as photosynthetic radiation are used to estimate NPP and, finally, NPP is used to estimate EP. The first step, estimation of chlorophyll, is known to have significant bias (underestimation by ~2-3 times) in the Southern Ocean which is transferred to higher level products. We correct for that bias by using algorithms tuned to Southern Ocean datasets blended with more or less standard products elsewhere [Mitchell and Kahru, 2009; Kahru and Mitchell, 2010]. While our satellite estimates of EP are improved, they are still subject to high uncertainty.

Observed seasonal cycles in APO provide a new benchmark for the ocean biogeochemistry model components of ESMs. They offer evaluation metrics complementary to ocean color products by providing additional information on deep ventilation processes unavailable from satellite data alone. The main drawback of APO seasonal cycles is that atmospheric transport models (ATMs) are needed to translate ocean model air-sea O_2 fluxes into a seasonal APO signal, which inevitably introduces uncertainty [*Stephens et al.*, 1998; *Nevison et al.*, 2012a]. A first attempt has been made to use APO seasonal cycles to evaluate ocean-only marine biogeochemistry models [*Naegler et al*, 2007], but the models in that study implemented a simplified parameterization of the biological processes affecting O_2 and CO_2 air-sea fluxes and were considerably less advanced than the current ecosystem dynamics and biogeochemical components used in state-of-the-art ESMs. Further, while Naegler *et al.* asserted that the uncertainty introduced by ATMs was too large to provide a strong constraint on ocean model fluxes, their study relied on only two ATMs. Here, we translate the model air-sea fluxes into APO signals using a wider range of ATMs and show that, in many cases, the discrepancies between modeled and observed APO seasonal cycles transcend ATM uncertainty.

2. Methods

2.1 Ocean Biogeochemistry Models

The CMIP5 models analyzed in this study include the Geophysical Fluid Dynamics Laboratory (GFDL) Earth System Models (depth-based ESM2M and density-based ESM2G vertical oceans; *Dunne et al.*, 2012) from Princeton, New Jersey; the Institut Pierre-Simon Laplace Coupled Model 5 in its low resolution version (IPSL-CM5A-LR, referred to as IPSL in the following) model from Paris, France; the Community Ecosystem Model (CESM) from the National Center for Atmospheric Research in Boulder; CO, the Max Planck Institut fuer Meteorologie (MPIM) Earth System Model, version MPI-ESM-LR, from Hamburg, Germany; and the Norwegian Earth

System Model (NorESM1-ME, referred here as NorESM1). The ocean biogeochemical models embedded in the respective ESMs are represented by TOPAZ (GFDL) [*Dunne et al*, 2013], PISCES (IPSL) [*Aumont and Bopp* 2006], BEC (CESM) [*Moore et al.,* 2002, 2004, 2013], and HAMOCC (MPIM) [*Ilyna et al.*, 2013]. NorESM1 uses a variant of HAMOCC, adapted to a sigma coordinate ocean circulation model [*Tjiputra et al.,* 2013].

The six ESMs differ in their physical components and implement ocean biogeochemical schemes that vary in their specifics, but have many common features. All include explicit representations of upper ecosystem dynamics that distinguish at least one phytoplankton group and one size class of zooplankton. Four of the models (CESM, both GFDL variants and IPSL) divide phytoplankton further into at least 2 size classes: large (micro) and small (nano + pico). GFDL and CESM also explicitly model diazotrophs. Phytoplankton growth rates in all models are co-limited by light, temperature and nutrient (N, P, Si, Fe) availability. Carbon export flux is closely linked to ecosystem structure and dynamics, with higher sinking rates assumed for large phytoplankton, representing, e.g., diatoms.

For each model, the following output fields were obtained for the CMIP5 standard historical simulation, which is driven by prescribed atmospheric CO_2 from 1850-2005: carbon export flux at 100 m depth (EP_{100}), vertically integrated NPP, net air-sea O_2 and CO_2 fluxes, net surface heat flux (Q), and sea surface salinity and temperature (SST). Many of these fields were available through public web interfaces, but some variables, particularly Q, required assistance from the individual modeling groups, which effectively limited the study to 6 models listed above. The EP_{100} and NPP fields were compared directly to the corresponding satellite ocean color products. The remaining 5 output fields were used in the estimation of APO time series, with the final three fields used to estimate air-sea N_2 fluxes based on the $Q(dS/dT)_{N2}/C_p$ equation [*Keeling et* *al.*, 1993; *Manizza et al.*, 2012] with modifications from *Jin et al.* [2007]. In this equation, Q is heat flux, $(dS/dT)_{N2}$ is the temperature derivative of the N_2 solubility coefficient, and C_p is the heat capacity of sea water. The resulting N_2 fluxes, together with the prognostic O_2 and CO_2 air-sea fluxes, were used as described below to force atmospheric transport model simulations to compute atmospheric time series of APO [*Naegler et al.*, 2007; *Nevison et al.*, 2008; 2012a]. Since all the ocean models operated on an irregular, off-polar grid with 2-dimensional latitude and longitude coordinates, these were first interpolated to a regular 1°x1° latitude/longitude grid using Climate Data Operators freeware (https://code.zmaw.de/projects/cdo). The CDO interpolation was not mass conservative, but resulted in global O_2 flux differences generally of less than 1%. An exception was the CESM, whose output was converted conservatively to a regular grid using a CESM-specific mapping file.

2.2 Atmospheric Transport Model Simulations

2.2.1 Matrix Method

A matrix method was used to translate the ocean model air-sea O_2, N_2 and CO_2 fluxes into corresponding annual mean cycles in atmospheric potential oxygen (APO). The method was based on the pulse-response functions from the Transcom 3 Level 2 (T3L2) atmospheric tracer transport model (ATM) intercomparison. Each of the 13 ATMs that participated in T3L2 conducted forward simulations in which a uniformly distributed CO_2 flux, normalized to 1PgC yr^{-1}, was released from each of 11 ocean regions (Figure 1) for each of 12 "emission months," i.e., Jan-Dec, allowed to decay for 35 months, using an annually repeating cycle of meteorology that was model specific for each ATM, and sampled every month at a range of surface monitoring sites [*Gurney et al.*, 2003; 2004]. The APO code was developed from an earlier pulse-response matrix code, which has been described in detail in *Nevison et al.* [2012b], that translates terrestrial net ecosystem exchange (NEE) fluxes of carbon into the corresponding annual mean cycles in atmospheric CO_2. The matrix method is considerably faster than a full forward ATM simulation, allowing annual mean cycles in APO from 13 different ATMs to be computed in seconds, rather than the days or weeks required for a single forward simulation.

The pulse-response matrix code was applied separately to the O_2, N_2 and oceanic CO_2 fluxes from the last 12 years of the historical simulations, spanning 1994-2005, converting from carbon to oxygen or nitrogen units where appropriate, to create three separate time series of atmospheric O_2, N_2 and CO_2 as mole fraction anomalies (μmol mol^{-1}) on a H_2O-free basis, where the O_2 and N_2 anomalies are computed as though O_2 and N_2 were trace gases, similar to CO_2. These were combined to calculate a 9-year time series in APO in per meg units, spanning fluxes from 1997-2005, according to Equation 1 [Stephens et al., 1998]:

$$APO = \frac{1}{X_{O2}}(O_2) - \frac{1}{X_{N2}}(N_2) + \frac{1.1}{X_{O2}}(CO_2), \qquad (1)$$

where X_{O2} and X_{N2} are the dry air mole fractions of O_2 and N_2 in H_2O-free air, treated here as constants (0.2094 and 0.7808, respectively). The mean seasonal cycle was computed by detrending the time series with a 3rd order polynomial and then taking the average of the detrended data for all Januaries, Februaries, etc. The matrix method involves calculating separately the components of APO at each measurement site arising from fluxes from each ocean region. These components are then summed to compute the net APO signal. The model definition of APO in Equation 1 ignores contributions to APO from land biospheric exchanges at ratios other than 1.1 and fossil fuel burning, but these are very small in comparison to oceanic contributions on seasonal time scales [*Manning and Keeling*, 2006; *Nevison et al.*, 2008].

2.2.2 Evaluation of matrix method based on APO Transcom

An evaluation exercise was conducted in which the APO pulse-response matrix code was forced by climatological O_2 and N_2 fluxes from *Garcia and Keeling* [2001] and used to compute the mean seasonal cycle in APO as described above using Equation 1 (minus the oceanic CO_2 term). The matrix-based results were evaluated against the mean seasonal cycles from archived station output from the forward ATM simulations of the APO Transcom Experiment, which also used the Garcia and Keeling O_2 and N_2 forcing fluxes [*Blaine*, 2005; *Nevison et al.*, 2012b]. This evaluation was conducted using a subset of 9 of the original 13 T3L2 ATMs that also participated in APO Transcom. For this subset, the matrix method performed well in relatively homogeneous regions like the Southern Ocean and at northern high latitude sites like Barrow, Alaska (BRW) and Alert, Canada (ALT). It was less reliable in capturing the forward simulation cycle at sites located within Northern midlatitude ocean regions, including Cold Bay, Alaska and La Jolla, California, where the uniform distribution of fluxes assumed by T3L2 did not accurately capture the impact of strong heterogeneity in air-sea fluxes from these regions (Supplementary Tables S1, S2 and Supplementary Figures S1, S2). These same North Pacific stations are subject to large uncertainty in full forward ATM simulations due to uncertainty in vertical mixing [*Stephens et al.*, 1998; *Battle et al*., 2006; *Tohjima et al.*, 2012]. We therefore focus in Section 3 on ALT, BRW and three Southern Ocean sites, including Macquarie Island (MQA), Palmer Station, Antarctica (PSA) and South Pole (SPO) in our use of APO to evaluate the ESM-simulated air-sea O_2, N_2 and CO_2 fluxes. The locations of these 5 sites with respect to the T3L2 ocean regions is shown in Figure 1.

While the evaluation exercise indicates that the matrix method reproduces the shape and phase of the seasonal cycles with high reliability at the above sites, it tends to underestimate the seasonal amplitude by about 4-5% at ALT and BRW and by 11-12% at MQA and SPO and to slightly overestimate the amplitude at PSA. In applying the matrix code to the ESM oceanic fluxes, we therefore scaled up the estimated cycles by site and ATM-specific scaling factors obtained from the evaluation exercise (Supplemental Tables S1, S2, Supplemental Figure S2). Since these scaling factors were only available for the subset of 9 of the 13 T3L2 ATMs that also participated in APO Transcom, we subsequently (Section 3.1) compare observations alternatively to the scaled 9-model subset, or to all 13 unscaled models.

2.2.3 Component O_2 Fluxes

The net air-sea O_2 flux for each ESM can be divided into three components, associated with NCP, deep ventilation and thermal processes [*Nevison et al.*, 2012a]:

$$F_{O2,total} = F_{O2,NCP} + F_{O2,vent} + F_{O2,therm} \qquad (2)$$

These in turn can be used to force the matrix model and the resulting total APO cycle can be presented as the sum of component signals according to Equation 3.

$$APO = APO_{NCP} + APO_{vent} + APO_{therm} \qquad (3)$$

Here, the APO_{therm} term also includes the effects of N_2 fluxes, as per the second right-hand term in Equation 1. The atmospheric signal due to oceanic CO_2 (last term in Equation 1) is not easily included in any of the component terms in Equation 3 based on available ESM output, but in principle all three component processes may lead to changes in CO_2 fluxes as well as O_2 fluxes. In practice, CO_2 has only a small influence on the amplitude and phasing of APO in most of the ESMs and thus is ignored in the component analysis. An exception is MPIM, in which the oceanic CO_2 signal has a peak-to-trough seasonal amplitude of up to 5 ppm in the Southern Ocean that opposes the O_2 cycle, as noted previously [*Anav et al.*, 2013] and discussed further below.

Among the terms in Equation 2, $F_{O2,total}$ was provided outright by the ESMs and the thermal component $F_{O2,therm}$ can be derived easily from standard ESM output following the approach described above for N_2. The remaining terms, $F_{O2,NCP}$ and $F_{O2,vent}$ are more challenging to estimate from available ESM output. In *Nevison et al.* [2012a], $F_{O2,NCP}$ was estimated from EP multiplied by a molar ratio of 1.4 mol O_2 per mol C exported. The assumption that $F_{O2,NCP} = 1.4$ EP was shown in *Nevison et al.* [2012a] to yield reasonable results for EP derived from satellite data (and indeed was applied to the satellite data described below in Section 2.3), but this approach proved unsatisfactory for EP_{100} from the ESMs, especially in the Southern Ocean as discussed further below, since it yielded an atmospheric signal that was unreasonably small. The assumption also led to phasing uncertainties for some models (IPSL, NorESM1 and MPIM) that use finite sinking velocities for particulate organic carbon (as opposed to instantaneous vertical redistribution, as assumed, e.g., by CESM) with a resulting delay in EP_{100} relative to NPP. Since the timing of $F_{O2,NCP}$ is likely to be more closely related to NPP than EP_{100} [*Nevison et al.*, 2012a], we estimated $F_{O2,NCP}$ from the ESMs alternatively as $1.4EP_{100}$ and 1.4 *ef* * NPP, where NPP is the standard, vertically-integrated ESM output variable and *ef* is the model-specific annual mean EP_{100}/NPP ratio, integrated over the 40-60°N or 40-60°S latitude band for northern and southern stations, respectively (Table 1).

Finally, $F_{O2,vent}$ in principle can be estimated as a residual of the other 3 terms in Equation 2. $F_{O2,vent}$ was estimated with reasonable success at the northern hemisphere sites, but generally looked unreasonable in the Southern Ocean for most models, with the exception of IPSL. The signals were judged to be unreasonable on the basis of whether the APO_{vent} term, if estimated as a residual from Equation 3, dominated the APO_{NCP} term in driving the springtime rise in APO. In reality, the APO_{NCP} term must be primarily responsible for this rise [*Keeling et al.*, 1993; *Bender et al.*, 1996; *Nevison et al.*, 2012a]. We therefore do not attempt to explicitly resolve or present APO_{vent} signals in the Southern Hemisphere. While the problems with APO_{vent} necessarily imply a corresponding problem in one or both of the other component terms APO_{NCP} and APO_{therm}, as discussed below, the shape of these latter terms is still informative and is less sensitive to the uncertainties inherent in the residually-estimated APO_{vent} term.

2.3 Satellite Ocean Color Data

The primary output product of satellite ocean color measurements historically has been the concentration of chlorophyll-a (Chl), which is also the main input to most satellite-based ocean primary productivity models [*Behrenfeld and Falkowski*, 1997]. However, the standard Chl product based on empirical band-ratios of reflectances represents primarily the coefficient of total absorption of blue light and is inherently biased if the distributions of the optically active components deviate from the global "mean" [*Lee et al*., 2011; *Siegel et al*., 2005; *Sauer et al*., 2012]. In the Southern Ocean the standard Chl algorithms underestimate *in situ* Chl by 2-3 times [*Mitchell & Kahru*, 2009] whereas in the Arctic they overestimate it [*Mitchell*, 1992]. These errors are directly transferred into errors in estimates of net primary production (NPP) and export production (EP).

For the Southern Hemisphere we used an empirical Chl algorithm (SPGANT) that was tuned to in situ Chl in the Southern Ocean and spatially blended with the standard SeaWiFS OC4 algorithm [*Kahru and Mitchell*, 2010]. The same blending scheme was applied when blending NPP between two versions of the Vertically Generalized Productivity Model (VGPM) algorithm [*Behrenfeld & Falkowski*, 1997]: the Southern Ocean version and the low-latitude version of *Kahru et al*. [2009]. EP was calculated using a modified version of the *Laws* [2004] model according to *Nevison et al*. [2012a]. The mean annual cycles for Chl, NPP and EP were calculated for 1997-2009 using data derived from SeaWiFS.

For the Northern Hemisphere we used NPP data calculated according to the standard VGPM using MODIS-Aqua Chl. NPP was downloaded from http://science.oregonstate.edu/ocean.productivity. EP was calculated according to *Dunne et al.* [2005]. The mean annual cycles for NPP and EP were calculated for 2002-2011 using monthly composites derived from MODIS-Aqua. While the *Laws* [2004] and *Dunne et al.* [2005] methods of deriving EP are not identical, they both estimate export efficiency as a function of sea-surface temperature and NPP, are fitted to *in situ* data, and generally produce similar estimates. In *Nevison et al.* [2012a] the Southern Ocean EP derived with the Laws model was modified by constraining to the bulk nutrient budget estimated in the ocean inversion of *Schlitzer* [2000]. That reduced the unrealistically high export efficiency of the Laws model observed at cold temperatures and brought it into closer agreement with the Dunne *et al.* export efficiency.

Both the SPGANT and VGPM/OSU satellite algorithms for NCP were converted to air-sea O_2 fluxes using $F_{O2,NCP} = 1.4$ NCP, where 1.4 refers to the molar ratio between O_2 produced and carbon fixed in photosynthesis. $F_{O2,NCP}$ was used to force the pulse-response code to estimate the corresponding APO signal associated with NCP as per *Nevison et al.* [2012a].

2.4 APO Data

APO is a unique atmospheric tracer of ocean biogeochemistry that is calculated by combining high precision O_2 and CO_2 data according to APO = O_2 + 1.1CO_2 [*Stephens et al.,* 1998]. By design, APO is mostly insensitive to exchanges with the land biosphere, which have a nearly fixed stoichiometry that produces compensating changes in O_2 and CO_2. In contrast, the exchanges of O_2 and CO_2 across the air-sea interface are not strongly correlated, largely because variability in dissolved CO_2 is strongly damped by carbonate chemistry in seawater on seasonal timescales. As a result, seasonal variability in APO reflects changes in atmospheric oxygen occurring almost solely due to oceanic processes [*Manning and Keeling*, 2006].

Atmospheric O_2 data, reported in terms of deviations in the O_2/N_2 ratio, were obtained from the Scripps Institution of Oceanography (SIO) and Princeton University (PU) networks. Data are available from the early to mid 1990s, depending on the station [*Bender et al.*, 2005; *Manning and Keeling*, 2006]. In Figure 2, we use SIO data from SPO, PSA and ALT and PU data from MQA and BRW. Details of the station locations and time spans of data used to calculate the mean seasonal cycle are listed in Table S2 and shown in Figure 1. For MQA (1997-2007) and BRW (1993-2008), the time spans overlapped mostly but not perfectly with the CMIP5 model output (1994-2005) and the satellite data (1997-2009 for SPGANT, 2002-2011 for VGPM).

APO was calculated according to,

$$\mathrm{APO} = \delta(O_2/N_2) + \frac{1.1}{X_{O2}} CO_2, \qquad (4)$$

where $\delta(O_2/N_2)$ is the relative deviation in the O_2/N_2 ratio from a reference ratio in per meg units, X_{O2} = 0.2094 is the O_2 mole fraction of dry air [*Tohjima et al.,* 2005], CO_2 is the mole fraction of carbon dioxide in parts per million (μmol mol^{-1}), and 1.1 is a qualitative estimate of the -O_2:C ratio of terrestrial respiration and photosynthesis. Mean seasonal cycles for observed APO were obtained using the same detrending and averaging methodology described in Section 2.2.1. The uncertainty in the observed mean seasonal cycles over the timespan of available data is less than 6% at extratropical latitudes, reflecting a combination of instrumental precision, synoptic variability and interannual variability (IAV) in the seasonal cycle. The current study is focused on the mean seasonal cycle.in APO as a first order challenge for the CMIP5 ocean models. Here, model, APO and satellite seasonal cycles are evaluated over roughly comparable periods that are dictated by data availability. The examination of interannual variability is deferred to future research, which will require ATM simulations of APO driven by interannually varying meteorology.

2.5 Phase Metrics

The time of year of the seasonal maximum in APO and NPP was used as a phase metric. For APO, monthly mean, station-specific time series, both modeled and observed, were fit to a 3rd order polynomial plus first 2 harmonics function. The harmonic components of the fit were used to construct a mean seasonal cycle with daily resolution and the day of the seasonal maximum was identified. The same approach was used to derive the day of the seasonal NPP maximum, except that the fit was applied to monthly mean satellite-derived and ESM NPP integrals summed from 40-60°S and 40-60°N, which were compared to the APO phase metric at southern and northern stations, respectively.

3. Results

3.1 APO comparison to Earth System Models

The APO cycles estimated from the 6 sets of ESM air-sea fluxes were compared to observations at 3 Southern Ocean and 2 northern monitoring sites (Figure 2). In these plots, the green envelope reflects our best estimate of the ATM uncertainty in the ocean model APO signal based on the 9 scaled ATM results, while the gray window reflects the more complete range of uncertainty using all 13 unscaled ATM results. In general, the distinction between the green and gray windows is only moderately important, as the observed APO cycle in most cases either falls within both envelopes or lies outside of both envelopes.

The MPIM and related NorESM1 ocean biogeochemistry models are examples in which the observed APO cycle lies outside both ranges of uncertainty at all 5 evaluation sites (Figure 2, lower middle and right panels). For these models, the rise in the APO cycle occurs too early in the springtime in both hemispheres, while the overall amplitude of the cycle is too large at all the southern stations. Here, it is notable that the MPIM APO amplitude would be even larger in the Southern Ocean if it were not offset by the unrealistically large seasonal cycle in oceanic CO_2 described above. The large CO_2 cycle, however, does not substantially alter the phase of APO, which is determined mainly by the timing of the O_2 fluxes.

IPSL is another ocean biogeochemistry model for which the observed APO cycle lies outside of both the best guess and full range of uncertainty at all monitoring sites, with the exception of Palmer Station (64.9°S), where observed APO falls within the wider gray window of uncertainty (Figure 2b, lower left panels). Unlike MPIM and NorESM1, the rise in the IPSL APO cycle occurs somewhat later in the springtime than observed, while the overall amplitude of the cycle tends to be underestimated. The underestimate is mild at all the southern stations, and even falls within the broader range of uncertainty at PSA, but is more pronounced at the northern monitoring sites, where the IPSL amplitude is too small by nearly a factor of 2.

CESM is the top-performer among the 6 ESMs evaluated, consistently yielding green (gray) windows that encompass the observed APO cycle at most (all) of the 5 monitoring sites (Figure 2, upper left panels). GFDL ESM2M (depth-based coordinates) is the second most consistent performer, yielding cycles that generally agree with observations, with exceptions at BRW, where ESM2M tends to mildly underestimate the depth of the APO trough, and at PSA, where the rise in the APO cycle may be up to 1 month too early. The sigma-coordinate GFDL ESM2G model is the third best performer, capturing the observed APO cycle relatively well at most southern stations, but underestimating the seasonal amplitude at the northern stations.

3.1.1 Regional analysis of APO cycle

The matrix method can partition the ocean model APO cycles into regional contributions from the 11 ocean regions used in T3L2. At the southern stations of SPO, PSA, and MQA, this partitioning reveals, not surprisingly, that the Southern Ocean (defined as all ocean regions south of 44°S) dominates the APO cycle (not shown). However, at BRW and ALT at least 3 regions make important contributions, including the "temperate" North Pacific (extending from 15°N to the Bering Strait around 65°N and thus including the subpolar region), the "temperate" North Atlantic (extending from 15°N to 48°N) and the "Northern Ocean" (including the Arctic Ocean and the North Atlantic north of 48°N). The Northern Ocean is the most important contributor to the APO seasonal cycle at both BRW (Figure 3) and ALT and is by far the most variable component among the 6 ESMs. The largest Northern Ocean APO amplitudes are produced by CESM and NorESM1, which are the only two models that capture the total observed APO amplitude at BRW (Figure 2d).

3.1.2 Partitioning of APO cycle among component signals

To probe further into the underestimate of the APO amplitude at BRW by most of the ESMs, we partitioned APO into thermal and NCP-related components, as described in Section 2.2.3 (Figure 4). A comparison of CESM and ESM2M in Figure 4 indicates that both have similar APO_{therm} and APO_{NCP} signals, but that CESM captures total APO more or less correctly while ESM2M underestimates the total APO amplitude. By inference, the missing APO_{vent} term accounts for the difference. However, as discussed in Section 2.2.3, APO_{vent} can be estimated only as a residual of 3 other terms using standard CMIP5 output and thus its shape and phasing are sensitive to even small uncertainties in those other terms. Thus, the residual ventilation curves in Figure 4 should be interpreted with caution (e.g., the MPIM curve is clearly unreasonable in phasing). The four remaining ESMs have APO_{NCP} cycles of similar or smaller amplitude than CESM, which in the case of ESM2G and MPIM is due primarily to their relatively low ef-ratios, and all these models substantially underestimate the total APO amplitude at BRW. This suggests that these models probably also underestimate some combination of deep ventilation and NCP.

A similar partitioning of APO was attempted in the Southern Ocean, but the estimation of APO_{NCP} from model EP_{100} generally did not give plausible results in this region. This problem is discussed in more detail in Section 4.

3.2 Satellite data compared to ESMs

Estimates of net primary production display a wide variety of spatial patterns among models and satellite data (Figure 5). Global totals range over more than a factor of 2 (34-82 Pg C/yr) among the ESMs, with most models tending to exceed the VGPM satellite-based estimate of 45 Pg C/yr (Table 1). Global EP is more consistent among the models, with a value around 8 Pg C/yr in most cases, in good agreement with the satellite-based estimate. Global EP converges among the ESMs because the model with highest global NPP (ESM2M) has a small ef-ratio of < 0.1 and the models with lowest global NPP (IPSL, NorESM1) have the largest ef-ratios of about 0.2 (Table 1).

The high global NPP totals in the ESMs are driven in large part by high tropical NPP values, which generally are not reflected in the satellite data except along coastlines (Figure 5). In this paper, we focus on the 40-60° latitude bands, which are more important than the tropics in driving the seasonal cycles in NPP, EP (NCP) and APO [*Garcia and Keeling*, 2001; *Anav et al.*, 2013]. In the Southern Ocean 40-60°S band, global NPP ranges among ESMs from 5.2 to 12.5 PgC/yr, encompassing the satellite-based estimates (Table 1, Figure 6). However, the ESMs tend to underestimate EP relative to the satellite-derived values, particularly the SPGANT/Laws product, due largely to the small model ef-ratios. In the 40-60°N band, the ESMs generally underestimate both NPP and ef-ratios relative to the satellite-derived values. This combination leads to model EP values that are smaller than satellite EP by a factor of 2 on average (Table 1). In both hemispheres, the model NPP maximum tends to occur earlier than the satellite-derived maximum, with some models (IPSL, MPIM) predicting a maximum that is up to 1-2 months early (Figure 6).

3.3 Combining APO and Satellite Data

In the previous sections we considered APO and satellite data as separate evaluation metrics for ESMs. Below we consider the two as combined metrics. While this analysis is limited by uncertainties in the absolute magnitude of satellite NPP and EP/NCP and our imperfect ability to partition the ESM total APO signal into its NCP and other components, it nevertheless provides some additional insight into the behavior of the ESMs.

3.3.1 Phase metrics

The phase metrics defining the timing of the observed and model seasonal maximum in APO reveal characteristic patterns for each ESM, which are relatively consistent across APO monitoring sites (Figure 7). The APO seasonal maxima of MPIM and NorESM1 are earlier than observed by about 1 month and 3 weeks, respectively, on average, while the IPSL APO maximum (with the exception of PSA) tends to be later than observed by 2-3 weeks. The remaining models, CESM, ESM2M and ESM2G, have seasonal APO maxima that are relatively consistent with observations, although with some variation among different stations.

The observed seasonal maximum of NPP occurs about 30-40 days earlier than the observed APO maximum in the Southern Ocean stations and about 50 days earlier at BRW and ALT. Of the models, ESM2G, CESM and ESM2M capture the phase of the NPP maximum to within about 1-3 weeks, although as noted above in Figure 6 the model NPP maxima tend to occur earlier than the satellite-based maxima. In MPIM, the NPP maximum is about 1 to 1.5 months earlier than observed, and the APO maximum is also corresponding early (Figure 7). IPSL is an outlier from the general slope of the APO vs. NPP phase relationship, as defined by the rest of the ESMs. The IPSL NPP maximum occurs about 40 days earlier than observed in the Southern Hemisphere and nearly 2 months earlier than observed in the Northern Hemisphere, but IPSL, curiously, also has the latest APO seasonal maximum of any of the models. NorESM1 is another outlier in the opposite direction off the general APO vs. NPP phase slope, at least in the Northern Hemisphere. There, NorESM1's seasonal maximum in NPP has a relatively small lag from the APO maximum compared to the other models. NorESM1 is also unusual in that the APO_{therm} seasonal maximum at Barrow occurs about 1 month later than in any of the other ESMs (Figure 4).

3.3.2 Seasonal amplitudes

In addition to evaluating the phasing of the ocean model APO and NPP cycles, we examined the amplitude of the cycles, with the caveat that the absolute magnitude of satellite-based NPP is not well determined and at present provides a relatively weak constraint on the models. Furthermore, the APO seasonal amplitude in principle is more closely related to NCP (or EP) than NPP. However, we chose NPP for the seasonal amplitude analysis due to the strong discrepancies in ef-ratio among models and satellite data indicated in Table 1, which may unduly bias the results.

A cross plot of the seasonal amplitude in APO against the seasonal amplitude of NPP integrated between 40-60°S suggests a strong correlation between the amplitudes of APO and NPP among the ocean biogeochemistry models, with larger NPP amplitudes associated with larger APO cycles. The strong correlation holds at all Southern Ocean stations and is illustrated in Figure 8a at Macquarie. The cluster of top-performing ESMs (CESM, ESM2M, ESM2G) agrees relatively well with the observed APO and SPGANT amplitudes. Meanwhile both amplitudes are underestimated by IPSL and overestimated by NorESM1 and MPIM.

Cross plots of the seasonal amplitudes of APO and NPP in the northern hemisphere reveals that these amplitudes are positively correlated at BRW (Figure 8b) and ALT (not shown), although the correlation is weaker than in the Southern Hemisphere. CESM, ESM2G, ESM2M and MPIM all capture the satellite-based NPP seasonal amplitude relatively well, while both CESM and

NorESM1 capture the observed APO amplitude accurately. However, CESM is the only model that reproduces both the NPP and APO seasonal amplitudes well relative to the observations.

4. Discussion

4.1 Northern Ocean

Most ESMs tend to underestimate substantially the observed seasonal amplitude of APO at Barrow, Alaska. A combination of region-specific results (Figure 3) and O_2 component analysis (Figure 4) suggests that some combination of fall/winter deep ventilation and spring/summer export production in the Northern Ocean (defined to include the North Atlantic north of 48°N) in particular may be underestimated in many models. The combined analysis of the APO vs. NPP seasonal amplitudes (Figure 8b) supports these conclusions and suggests that, while several models may be capturing primary production well in the Northern Ocean, accurate representation of export production and deep ventilation is also important for reproducing the observed APO cycle. The inference from the APO component analysis in Figure 4 that the GFDL models may have weak ventilation in the North Atlantic appears to contradict the analysis of *Dunne et al.* [2012], who found robust NADW formation in both the ESM2M and ESM2G versions, but possibly could be reconciled if the biogeochemical gradients across which deep water formation acts are too weak..

We investigated several mechanisms that might explain the differences among models in the APO cycle at high northern latitudes, including subpolar heat transport and Arctic sea ice cover. Here, stronger northward heat transport should lead to more deep ventilation, while lower sea ice cover will permit more production and ventilation in the Arctic Ocean. Subdividing the Northern Ocean region into Arctic Ocean and North Atlantic components revealed that some models (IPSL and ESM2G) have a very small component (< 2 per meg) of APO seasonal amplitude coming from the Arctic Ocean alone (Figure 9). In ESM2G this may be related to the extensive winter sea ice cover, which exceeds the observed covered area reported by the National Snow and Ice Data Center (http://nsidc.org/data/seaice_index/archives.html) by about 2 x 10^6 km^2. However, sea ice cover is lower than observed in IPSL, suggesting the small Arctic APO component in that model is more related to general underestimate of primary and export production (e.g., as shown in Figures 6b and 8b). While it seems clear that the strong APO seasonality in CESM can be attributed in part to its high productivity and EP in the northern subpolar and polar regions (Figure 6 and Table 1), a full explanation for the underlying mechanisms of the CESM fidelity on APO compared to the other models is not readily apparent from surface-only data. This suggests the need for a more detailed exploration of ocean interior ventilation and biological response interactions outside the scope of the present work.

4.2 Southern Ocean

Compared to the Northern Hemisphere stations, the ESMs generally are more successful in the Southern Ocean in capturing the observed APO cycle (Figure 2). Within the range of ATM uncertainty, at least 3 models, CESM, ESM2M, ESM2G (and IPSL at Palmer Station), predict seasonal APO amplitudes in agreement with observations. Although the Southern Ocean APO amplitude in these models varies over as much as 20 per meg, we currently are not able to distinguish which of the underlying air-sea O_2 flux fields is the most realistic, due to the uncertainty associated with translating these fluxes into an atmospheric signal using TransCom3 era model responses to uniformly distributed regional fluxes. However, even with our current matrix method, the APO constraint is sufficiently robust to indicate that NorESM1 and MPIM substantially overestimate some combination of production and deep ventilation in the Southern

Ocean, while IPSL probably tends to underestimate these fluxes (Table 1, Figure 8a). Notably, the ESMs that reproduce APO the best in the Southern Ocean tend to predict a smaller net carbon uptake between 44-75°, and are in better agreement with independent estimates [*Lenton et al.*, 2013] of carbon uptake from ocean inversions and observed pCO_2 databases (Figure 10).

Reducing ATM uncertainty is a challenge that potentially can be addressed by using column-integrated APO signals from aircraft data [Wofsy et al., 2011], or conversely, by using vertical profiles to identify top-performing ATMs [*Stephens et al.*, 2007]. In addition, the spread in ATM results has been reduced substantially for CO_2 inversions using post-Transcom3-era ATMs [*Peylin et al.*, 2013], suggesting that ATM uncertainty also may be reduced for forward simulations of APO. If this is the case, then new forward simulations with several different modern-era ATMs may be sufficient to characterize ATM uncertainty. Alternatively, it may be valuable to continue with a matrix-based approach, using basis functions from many ATMs, but with redefined regional boundaries that are not defined based simply on latitude, as in T3L2 (Figure 1), but rather that correspond to the biogeography of major ocean regions [*Fay and McKinley*, 2014]. The definition of such basis functions could help extend the utility of the matrix approach to lower latitude APO monitoring sites and allow for the partitioning of the Southern Ocean into multiple regions defined around biogeochemical function, while still retaining the advantages of the matrix method, i.e., the ability to quickly and easily compare multiple ATMs forced with the same air-sea fluxes.

A second complication in the Southern Ocean analysis is that the EP_{100} values reported by the ESMs clearly are not directly comparable to satellite NCP(EP) data, particularly our SPGANT product, and thus can not be translated with confidence into air-sea O_2 fluxes associated with NCP. A likely problem is that the 100 m depth horizon used to compute EP may not be comparable across satellite algorithms and ocean biogeochemistry models. EP_{100} will underestimate the model's true NCP-related O_2 outgassing flux if organic matter is respired as it sinks from the actual model mixed layer depth to 100m depth [*Najjar et al.*, 2007]. It is also puzzling that the ef-ratios predicted by the ESMs (Table 1) appear to have decreased considerably in some cases relative to those reported for earlier versions of the same models [*Steinacher et al.*, 2010]. For example, the Southern Ocean ef-ratios for MPIM and IPSL in that earlier study were about 0.2 and 0.4, respectively, compared to 0.14 and 0.27, respectively, in the current study. The mean global ef-ratio for the 6 ESMs in the current study is only 0.14 and, even in the Southern Ocean, is only 0.17 on average, compared to satellite-based estimates of 0.18 globally and about 0.3 at high latitudes.

The small ef-ratios in the GFDL models (of less than 0.1 globally and only 0.10 to 0.13 in the Southern Ocean) appear consistent with the relatively deep summer MLDs in the Southern Ocean, which even at their minimum are often deeper than 100 m in both ESM2M and ESM2G [*Dunne et al.*, 2012]. In CESM the Southern Ocean summer mixed layer depths (MLDs) are generally shallower than 100 m and in many regions are only around 10-40 m deep [*Moore et al.*, 2013]. The shallower summer MLDs may contribute to CESM's larger ef-ratio of 0.18, although this ratio is still small compared to the satellite-based estimates. The small GFDL ef-ratios may also be related to an overvigorous picophytoplankton component wherein a prochloroccus-like form is capable of competing relatively well even in cold polar waters. Small picophytoplankton are more likely to be reoxidized and remineralized within the mixed layer, whereas larger, heavier microphytoplankton (e.g., diatoms) are more likely to be exported out of the oceanic mixed layer [*Uitz et al.*, 2010].

4.3 Phase relationships

While much of our analysis focuses on the seasonal amplitude of APO and NPP at mid to high latitudes, both of these metrics involve relatively large uncertainty. This derives from Transcom3-era uniform flux ATM uncertainty in the case of APO, while for NPP the uncertainty results from the lack of strong constraints on the absolute magnitude of the satellite fluxes. In contrast, we have relatively high confidence in the phasing of model APO, as represented by the matrix method (see Supplementary Information) and in NPP observationally derived from satellite data, based on the close correspondence in phasing between the SPGANT and VGPM algorithms. For these reasons, we used a phase metric, i.e., the timing of the seasonal maximum, to examine relationships between observed and model APO and NPP. As in the seasonal amplitude analysis, MPIM, NorESM1, and IPSL displayed phasing patterns that tended to deviate from observations and the other three top-performing models, albeit in diverging ways. A complete diagnosis of the model physics responsible for the phasing anomalies (e.g., IPSL's early NPP maxima and late APO maxima) described in Section 3.3.1 is beyond the scope of this paper. Here we note mainly that the phase metrics are a robust and relatively good indicator of overall model performance with respect to APO.

5. Summary

We have used measurements of the seasonal cycles in APO to challenge and test the ocean model components of 6 ESMs. The model/data comparison reveals that three of the ESMs tested reproduce the observed cycles reasonably well, within the range of ATM uncertainty, while three do not. ESM performance in general is more favorable in the Southern Hemisphere than in the Northern Hemisphere, where most models appear to underestimate the wintertime ventilation of O_2-depleted deepwater that drives the declining branch of the APO seasonal cycle and many may also underestimate both primary and export production, particularly at high northern latitudes. We used NPP and NCP(EP) products derived from satellite ocean color data as complementary constraints on the models in an effort to tighten the APO constraint, which reflects a combination of production and ventilation processes. However, while the satellite data provide relatively strong constraints with respect to phasing, they are more uncertain with respect to the absolute magnitudes of NPP and NCP(EP).

At least two primary uncertainties limit our ability to place stronger constraints on ocean model biogeochemistry based on currently available information from APO and satellite data: 1) The relatively large ATM uncertainty involved in translating air-sea O_2 fluxes into APO signals. 2) The uncertainty in how model EP_{100} relates to the true model $F_{O2,NCP}$ flux and how this relationship varies across models and satellite algorithms. The first of these, ATM uncertainty, is large, as quantified using our Transcom3-based matrix method. However, it probably has been overstated in previous analyses, which in some cases went so far as to suggest that APO does not provide a useful constraint on ocean model fluxes [e.g., *Naegler et al.*, 2007]. Further, ATM uncertainty could be reduced substantially in future work with modern ATMs and O_2-specific flux patterns, or with new regional boundaries defined based on ocean biogeography rather than simple latitude. Even within the limits of our current approach, we have shown that half of the 6 ESMs tested here produce APO cycles whose mismatch with observed APO clearly transcends ATM uncertainty, suggesting underlying deficiencies in those models' physics and biogeochemistry.

Improving the understanding of the relationship between model air-sea O_2 fluxes and quantities like NPP, NCP and EP is a more tractable problem that can be dissected with appropriate model diagnostics, e.g., as per *Manizza et al.* [2012]. In the current analysis, using standard CMIP5 output from 6 ocean biogeochemistry models, we encountered difficulties in relating F_{O2} to EP and NCP, which hindered our ability to diagnose the mechanisms responsible for model performance and to compare ESM-derived APO_{NCP} directly to satellite-based APO_{NCP} signals. Extending model-derived insights to satellite products likely will require a shift in emphasis from EP at an arbitrary reference depth to near-surface processes like NCP, which are more relevant for exchanges of O_2 and CO_2 at the air-sea interface and more directly related to upward radiances detected by satellites.

Acknowledgements The authors gratefully acknowledge the CMIP5 ocean modelers for providing the output that made this project possible. In particular, we thank Keith Lindsay, Laure Resplandy, and Cristoph Heinze for their assistance. We also thank Britt Stephens and an anonymous reviewer for their helpful comments, which much improved the manuscript. Lastly, we thank Michael Bender and Robert Mika for providing APO data and acknowledge support from NASA Ocean Biology and Biogeochemistry grant NNX11AL73G.

References

Anav, A., and Coauthors, 2013: Evaluating the Land and Ocean Components of the Global Carbon Cycle in the CMIP5 Earth System Models. *J. Climate*, **26**, 6801–6843.

Arora, Vivek K., and Coauthors, 2013: Carbon–Concentration and Carbon–Climate Feedbacks in CMIP5 Earth System Models. *J. Climate*, **26**, 5289–5314.

Aumont, O., E. Maier-Reimer, S. Blain, and P. Monfray (2003), An ecosystem model of the global ocean including Fe, Si, P colimitations, *Global Biogeochem. Cycles*, 17, 1060, doi:10.1029/2001GB001745

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

Battle, M., Mikaloff Fletcher, S., Bender, M.L., Keeling, R.F., Manning, A.C., Gruber, N., Tans, P.P., Hendricks, M.B., Ho, D.T., Simonds, C., Mika, R. and Paplawsky, B. (2006), Atmospheric potential oxygen: New observations and their implications for some atmospheric and oceanic models, Global Biogeochem. Cycles, 20, GB1010, doi:10.1029/2005GB002534, 2006.

Behrenfeld, M.J., and P.G. Falkowski (1997), Photosynthetic rates derived from satellite based chlorophyll concentration, *Limnol. Oceanogr*., 42, 1-20.

Bender, M., Ho, D., Hendricks, M.B., Mika, R., Battle, M.O., Tans, P., Conway, T.J., Sturtevant, P. B., and Cassar, N. (2005), Atmospheric O_2/N_2 changes, 1993-2002: Implications for the partitioning of fossil fuel CO_2 sequestration, Global Biogeochem. Cycles, 19, GB4017, doi:10.1029/2004GB002410.

Blaine, T., W. 2005. *Continuous Measurements of Atmospheric Ar/N_2 as a Tracer of Air-Sea Heat Flux: Models, Methods, and Data*. University of California, San Diego, La Jolla.

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

Boyd, P. W. , S. C. Doney (2003) The impact of Climate Change and Feedback Processes on the Ocean Carbon Cycle. Chapter 7 , Ocean Biogeochemistry, The role of the Ocean Carbon Cycle in Global Change, Michael J. R. Fasham Editor, Springer-Verlag, pp. 157-198

Cadule, P., P. Friedlingstein, L. Bopp, S. Sitch, C. D. Jones, P. Ciais, S. L. Piao, and P. Peylin (2010), Benchmarking coupled climate-carbon models against long-term atmospheric CO_2 measurements, Global Biogeochem. Cycles, 24, GB2016, doi:10.1029/2009GB003556.

Dunne, J. P., R. A. Armstrong, A. Gnanadesikan, and J. L. Sarmiento (2005), Empirical and mechanistic models for the particle export ratio, *Global Biogeochem. Cycles*, 19, GB4026, doi:10.1029/2004GB002390.

Dunne, J. P., et al., 2012: GFDL's ESM2 global coupled climate-carbon Earth System Models Part I: Physical formulation and baseline simulation characteristics. *Journal of Climate*, doi:10.1175/JCLI-D-11-00560.1.

Dunne, J. P., 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*, doi:10.1175/JCLI-D-12-00150.1.

Fay, A. R. and McKinley, G. A.: Global open-ocean biomes: mean and temporal variability, Earth Syst. Sci. Data, 6, 273-284, doi:10.5194/essd-6-273-2014, 2014.

Gruber, N., and Sarmiento, J. L. (2002), Large-scale biogeochemical-physical interactions in elemental cycles, in *The Sea*, Volume 12, Robinson, A.R., McCarthy, J.J. and Rothschild, B.J. (eds.), John Wiley & Sons, Inc., New York.

Gurney, K. R., R. Law, A.S. Denning, P. Rayner, D. Baker, P. Bousquet, L. Bruhwiler, Y.-H. Chen, P. Ciais, S. Fan, I. Fung, M. Gloor, M. Heimann, K. Higuchi, J. John, E. Kowalczyk, T. Maki, S. Maksyutov, P. Peylin, M. Prather, B. Pak, J. Sarmiento, S. Taguchi, T. Takahashi, C-W. Yuen (2003), Transcom 3 CO_2 inversion intercomparison:

1. Annual mean control results and sensitivity to transport and prior flux information, *Tellus, Ser. B, 55*, 555–579.

Gurney, K. R., R. Law, A.S. Denning, P. Rayner, B. Pak, D. Baker, P. Bousquet, L. Bruhwiler, Y.-H. Chen, P. Ciais, S. Fan, I. Fung, M. Heimann, J. John, 2004, T. Maki, S. Maksyutov, P. Peylin, M. Prather, S. Taguchi, (2004), Transcom 3 inversion intercomparison: Model mean results for the estimation of seasonal carbon sources and sinks, *Global Biogeochem. Cycles, 18*, GB1010, doi:10.1029/2003GB002111.

Henson, S.A., Sarmiento, J.L., Dunne, J.P., Bopp, L., Lima, I., Doney, S.C., John, J., Beaulieu, C., 2010. Detection of anthropogenic climate change in satellite records of ocean chlorophyll and productivity. Biogeosciences 7, 621-640.

Henson, S. A., R. Sanders, E. Madsen, P. J. Morris, F. Le Moigne, and G. D. Quartly (2011), A reduced estimate of the strength of the ocean's biological carbon pump, Geophys. Res. Lett., 38, L04606, doi:10.1029/2011GL046735.

Henson, S., Cole, H., Beaulieu, C., and Yool, A.: The impact of global warming on seasonality of ocean primary production, Biogeosciences, 10, 4357-4369, doi:10.5194/bg-10-4357-2013, 2013.

Ilyina, T., K. D. Six, J. Segschneider, E. Maier-Reimer, H. Li, and I. N u~nez-Riboni (2013), Global ocean biogeochemistry model HAMOCC: Model architecture and performance as component of the MPI-Earth system model in different CMIP5 experimental realizations, J. Adv. Model. Earth Syst., 5, 287–315, doi:10.1029/2012MS000178.

Irwin, A. J. and M. J. Oliver (2009), Are ocean deserts getting larger ?, *Geophys. Res. Lett.*, 36, L18609, doi:10.1029/2009GL039883

Jiang, ChuanLi, Sarah T. Gille, Janet Sprintall, Colm Sweeney, 2014: Drake Passage Oceanic pCO2: Evaluating CMIP5 Coupled Carbon–Climate Models Using in situ Observations. *J. Climate*, **27**, 76–100.

Jin, X., Najjar, R.G., Louanchi, F., and Doney, S.C. (2007), A modeling study of the seasonal oxygen budget of the global ocean. J. Geophys. Res., 112, C05017, doi:10.1029/2006JC003731.

Kahru, M., R. Kudela, M. Manzano-Sarabia and B.G. Mitchell (2009), Trends in primary production in the California Current detected with satellite data, *J. Geophys. Res*., 114, C02004, doi:10.1029/2008JC004979.

Kahru, M. and B.G. Mitchell (2010), Blending of ocean colour algorithms applied to the Southern Ocean, *Remote Sensing Letters*, 1: 2, 119-124, doi: 10.1080/01431160903547940.

Keeling, R. F., Najjar, R. P., Bender, M. L., and Tans, P. P. (1993), What atmospheric oxygen measurements can tell us about the global carbon cycle, Global Biogeochem. Cycles, 7, 37–67.

Keeling, R.F., Stephens, B.B., Najjar, R.G., Doney, S.C., Archer, D., and Heimann, M. (1998), Seasonal variations in the atmospheric O_2/N_2 ration in relation to the kinetics of air-sea gas exchange, Global Biogeochem. Cycles, 12, 141-163.

Keeling R. F. A. Koertzinger, N Gruber (2010) Ocean Deoxygenation in warming world, *Annu. Mar. Rev. Sci*, 2:463-493.

Khatiwala, S., Primeau, F. and Hall, T. (2009) Reconstruction of the history of anthropogenic CO_2 concentrations in the ocean, Nature, 462, 346-350)

Laws, E. A., Falkowski, P. G., Smith, W. O., Jr., Ducklow, H. W., & McCarthy, J. J. 2000. Temperature effects on export production in the open ocean. *Global Biogeochemical Cycles, 14(4)*: 1231-1246.

Laws, E. A. (2004), Export flux and stability as regulators of community composition in pelagic marine biological communities: Implications for regime shifts, *Progress in Oceanography*, **60**(2-4), 343-353.

Lee, Z., et al. (2011), An assessment of optical properties and primary production derived from remote sensing in the Southern Ocean (SO GasEx), *J. Geophys. Res*., 116, C00F03, doi: 10.1029/2010JC006747.

Lenton., A., B. Tilbrook, R. Law, D. Bakker, S. Doney, N. Gruber, M. Ishii, M. Hoppema, N. Lovenduski, R. Matear, B. McNeil, N. Metzl, S. Mikaloff Fletcher, P. Monteiro, C. Roedenbeck, C. Sweeney, T. Takahashi (2013), Sea-air CO_2 fluxes in the Southern Ocean for the period 1990-2009, Biogeosciences, 10, 4037-4054.

Le Quéré, C., Roedenbeck, C., Buitenhuis ET, Conway TJ, Langenfelds R, et al. (2007), Saturation of the Southern Ocean CO2 sink due to recent climate change. Science 316, 1735-1738.

Maier-Reimer, E., Kriest, I., Segschneider, J., and Wetzel, P.: The HAMburg Ocean Carbon Cycle model HAMOCC5.1, Berichte zur Erdsystemforschung 14/2005, Max Planck-Institut fu ̈r Mete- orologie, Hamburg, Germany, 2005.

Manizza, M., R. F. Keeling , and C. D. Nevison, On the processes controlling the seasonal cycles of the air-sea fluxes of O_2 and N_2O: A modeling study, *Tellus-B*, 64, 18429, http://dx.doi.org/10.3402/tellusb.v64i0.18429, 2012.

Manning, A. C. , R. F Keeling (2006) Global oceanic and land biotic carbon sinks from the Scripps atmospheric oxygen flask sampling network, *Tellus-B*, 58B, 95-116

Mitchell, B.G. (1992) Predictive bio-optical relationships for polar oceans and marginal ice zones. *J. of Marine Systems*, 3: 91-105.

Mitchell, B.G. and M. Kahru (2009), Bio-optical algorithms for ADEOS-2 GLI, *J. Remote Sensing Soc. of Japan*, 29, 1, 80-85.

Moore, J.K., Doney, S.C.,Kleypas, J.C., Glover,D.M., Fung, I.Y., (2002) An intermediate complexity marine ecosystem model for the global domain, *Deep-Sea Res*. II, 49: 403-462.

Moore, J.K., Doney, S.C., and Lindsay, K. (2004) Upper ocean ecosystem dynamics and iron cycling in a global 3D model, *Global Biogeochem. Cycles*, 18, GB4028, doi:10.1029/2004GB002220

Moore, J., K. Lindsay, S. Doney, M. Long, and K. Misumi, 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. J. Climate, 26, 9291-9312, doi:10.1175/JCLI-D-12-00566.1.

Naegler T, P. Ciais, J. Orr, O. Aumont, C. Roedenbeck (2007) On evaluating ocean models with atmospheric potential oxygen, *Tellus*, Vol 59B, 138-156.

Najjar, R.G., et al. (2007) Impact of circulation on export production, dissolved organic matter, and dissolved oxygen in the ocean: Results from Phase II of the Ocean Carbon-cycle Model Intercomparison Project (OCMIP-2), Global Biogeochem. Cycles, 21, GB3007, doi:10.12007.

Nevison, C.D., R.F. Keeling, R.F. Weiss, B.N. Popp, X. Jin, P.J. Fraser, L.W. Porter and P.G. Hess (2005), Southern Ocean ventilation inferred from seasonal cycles of atmospheric N_2O and O_2/N_2 at Cape Grim, Tasmania, *Tellus*, 57B, 218-229.

Nevison, C. D., Mahowald, N. M., Doney, S. C., Lima, I. D., and Cassar, N.: Impact of variable air-sea O_2 and CO_2 fluxes on atmospheric potential oxygen (APO) and land-ocean carbon sink partitioning, Biogeosciences, 5, 875-889, 2008.

Nevison, C.D. E. Dlugokencky, G. Dutton, J.W. Elkins, P. Fraser, B. Hall, P.B. Krummel, R.L. Langenfelds, S. O'Doherty, R.G. Prinn, L.P. Steele, R.F. Weiss, Exploring causes of interannual variability in the seasonal cycles of tropospheric nitrous oxide, Atmospheric Chemistry and Physics, 11, doi:10.5194/acp-11-1-2011, 1-18, 2011.

Nevison, C. D., R. F. Keeling, M. Kahru, M. Manizza, B. G. Mitchell, and N. Cassar, Estimating net community production in the Southern Ocean based on atmospheric potential oxygen and satellite ocean color data, Global Biogeochem. Cycles, 26, GB1020, doi:10.1029/2011GB004040, 2012.

Nevison, C.D., D.F. Baker, and K.R. Gurney, A methodology for estimating seasonal cycles of atmospheric CO_2 resulting from terrestrial net ecosystem exchange (NEE) fluxes using the Transcom T3L2 pulse-response functions, *Geosci. Model Dev. Discuss., 5*, 2789-2809, 2012, www.geosci-model-dev-discuss.net/5/2789/2012/ doi:10.5194/gmdd-5-2789-2012, 2012b.

Peylin, P., R. Law, K. Gurney, F. Chevalier, A. Jacobson, T. Maki, Y. Niwa, P. Patra, W. Peters, P. Rayner, C. Roedenbeck, J. van der Laan-Luijkx, X. Zhang (2013), Global atmospheric carbon budget: results from an ensemble of atmospheric CO_2 inversions, Biogeosciences, 10, 6699-6720.

Polovina, J. J., E. A. Howell, and M. Abecassis (2008), Ocean's least productive waters are expanding, *Geophys. Res. Lett.*, 35, L03618, doi:10.1029/2007GL031745.

Roy, T., et al., 2011: Regional Impacts of Climate Change and Atmospheric CO2 on Future Ocean Carbon Uptake: A Multimodel Linear Feedback Analysis. *Journal of Climate*, **24**, 2300-2318.

Sauer, M.J., Roesler, C.S., Werdell, J.P., Barnard, A., 2012. Under the hood of satellite empirical chlorophyll algorithms: revealing the dependencies of maximum band ratio algorithms on inherent optical properties, Optics Express 20, 20920-20933.

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 and export production in three global coupled climate carbon cycle models, *Biogeosciences*, 5, 597-614.

Siegel, D. A., S. Maritorena, N. B. Nelson, and M. J. Behrenfeld (2005), Independence and interdependencies among global ocean color properties: Reassessing the bio-optical assumption, J. Geophys. Res., 110, C07011, doi:10.1029/2004JC002527.

Sigman, D.M., M.P. Hain, G.H. Haug, The polar ocean and glacial cycles in atmospheric CO_2, *Nature 466*, 47-55, 2010.

Six, K. and Maier-Reimer, E.: Effects of plankton dynamics on seasonal carbon fluxes in an ocean general circulation model, Global Biogeochem Cy., 10, 559–583, 1996.

Steinacher, M., F. Joos, T.L. Froelicher, L. Bopp, P. Cadule, V. Cocco, S.C. Doney, M. Gehlen, K. Lindsay, J.K. Moore, B. Schneider and J. Segschneider, Projected 21st centure decrease in marine productivity: a multi-model analysis, Biogeosciences, 7, 979-1005, 2010.

Stephens, B.B., Keeling, R.F., Heimann, M., Six, K.D., Murnane, R. and Caldeira, K. (1998), Testing global ocean carbon cycle models using measurements of atmospheric O_2 and CO_2 concentration, Global Biogeochem. Cycles, 12, 213-230.

Stephens, B.B., et al. (2007), Weak northern and strong tropical land carbon uptake from vertical profiles of atmospheric CO_2, *Science*, 316, 1732-1735, doi: 10.1126/science.1137004.

Tjiputra, J.F., C. Roelandt, M. Bentsen, D.M. Lawrence, T. Lorentzen, J. Schwinger, Ø. Seland, and C. Heinze, Evaluation of the carbon cycle components in the Norwegian Earth System Model (NorESM), Geosci. Model Dev., 6, 301-325, 2013

Tohjima, Y., H. Mukai, T. Machida, Y. Nojiri, and M. Gloor (2005), First measurements of the latitudinal atmospheric O2 and CO2 distributions across the western Pacific, Geophys. Res. Lett., 32, L17805, doi:10.1029/2005GL023311.

Tohjima, Y., C. Minejima, H. Mukai, T. Machida, H. Yamagishi, and Y. Nojiri (2012), Analysis of seasonality and annual mean distribution of atmospheric potential oxygen (APO) in the Pacific region, Global Biogeochem. Cycles, 26, GB004110, doi:10.1029/2011GB004110.

Uitz, J., H. Claustre, B. Gentili, and D. Stramski (2010), Phytoplankton class-specific primary production in the world's oceans: Seasonal and interannual variability from satellite observations, *Global Biogeochem. Cycles*, 24, GB3016, doi:10.1029/2009GB003680

Vancoppenolle, M., L. Bopp, G. Madec, J. Dunne, T. Ilyina, P. R. Halloran, and N. Steiner (2013), Future Arctic Ocean primary productivity from CMIP5 simulations: Uncertain outcome, but consistent mechanisms, Global Biogeochem. Cycles, 27, 605–619, doi:10.1002/gbc.20055.

Wofsy, S. C., the HIPPO Science Team and Cooperating Modellers and Satellite Teams: HIAPER Pole-to-Pole Observations (HIPPO): Fine grained, global scale measurements for determining rates for transport, surface emissions, and removal of climatically important atmospheric gases and aerosols, Phil. Trans. of the Royal Society A, 369(1943), doi:10.1098/rsta.2010.0313, 2073-2086, 2011.

Table 1: Vertically integrated NPP, EP at 100m (both in PgC yr-1) and EP/PP (ef-ratio) for 6 CMIP5 models and 2 Satellite Products.

Model	CESM	ESM2M	ESM2G	IPSL	NorESM1	MPIM	VGPM	SPGANT*
Global								
EP	7.97	7.78	5.27	7.02	8.00	8.26	8.20	N/A
NPP	56.3	82.2	66.5	33.6	41.0	57.9	45.42	N/A
EP/NPP	0.14	0.095	0.08	0.21	0.20	0.14	0.18	N/A
40-60N								
EP	0.71	0.83	0.53	0.75	0.66	0.51	1.47	N/A
NPP	3.85	4.71	3.92	2.42	3.45	3.77	4.97	N/A
EP/NPP	0.19	0.18	0.14	0.31	0.19	0.13	0.30	N/A
60-90N								
EP	0.34	0.33	0.21	0.19	0.15	0.08	0.46	N/A
NPP	1.48	1.35	0.95	0.58	0.74	0.75	1.29	N/A
EP/NPP	0.23	0.24	0.22	0.33	0.20	0.11	0.36	N/A
40-60S								
EP	1.25	1.18	0.82	1.42	1.93	1.77	1.60	2.85
NPP	6.77	9.36	8.53	5.24	10.3	12.5	6.01	8.81
EP/NPP	0.18	0.13	0.10	0.27	0.19	0.14	0.27	0.32

* SPGANT totals are only shown for the 40-60°S band because the algorithm is optimized for the Southern Ocean but not well validated in the Northern Hemisphere.

Figures

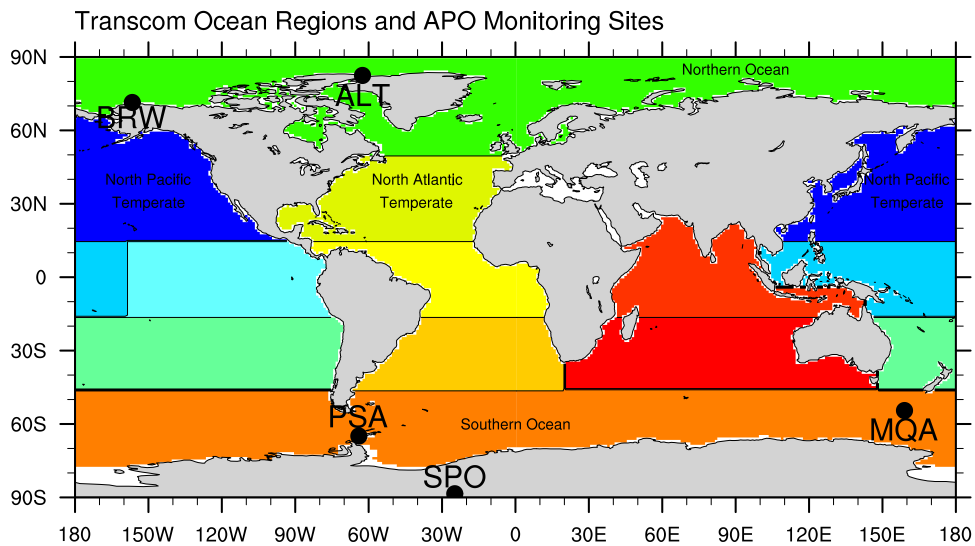


Figure 1. Transcom3 Level 2 ocean regions used in the matrix-based atmospheric transport method. Locations of the 5 APO monitoring sites featured in Figure 2 are superimposed.

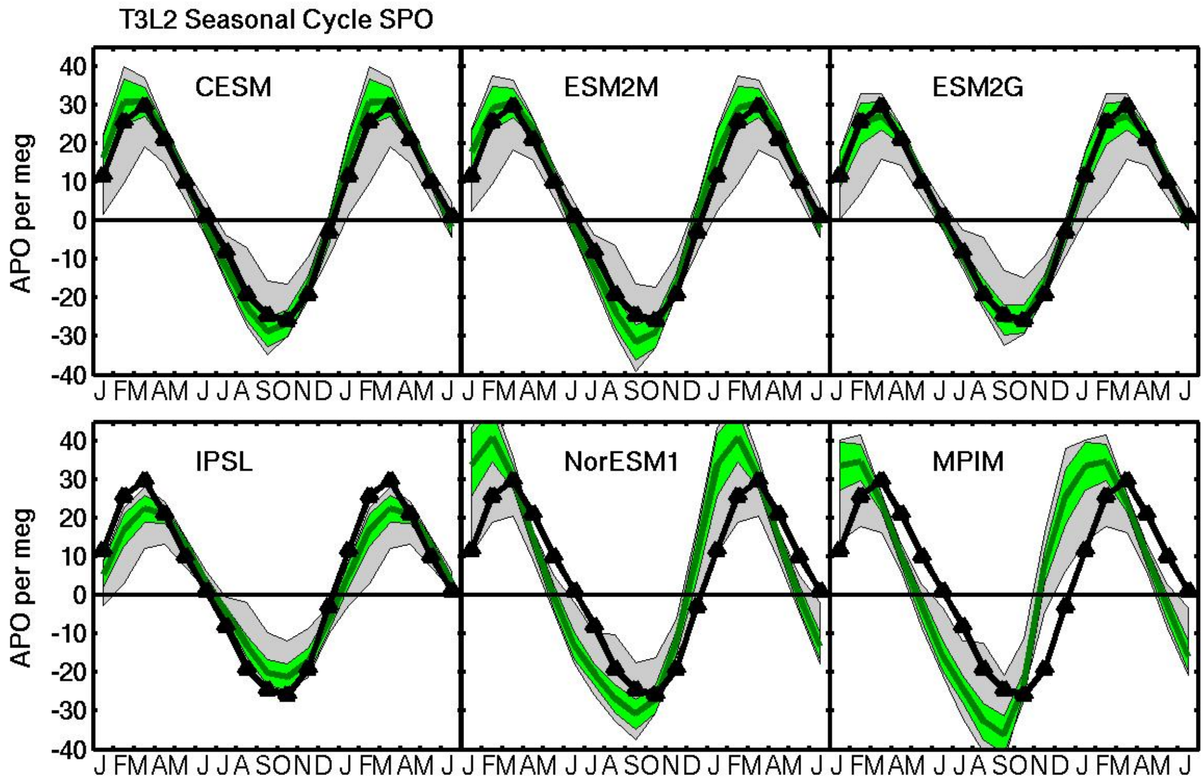
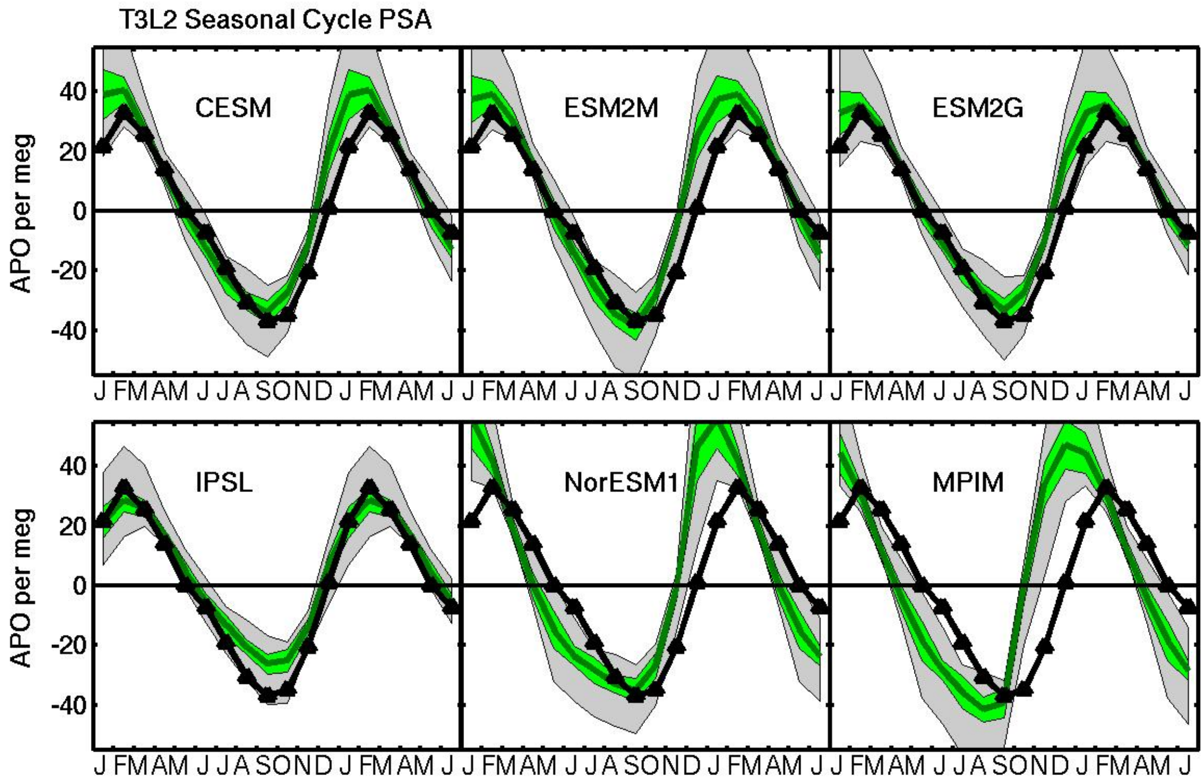
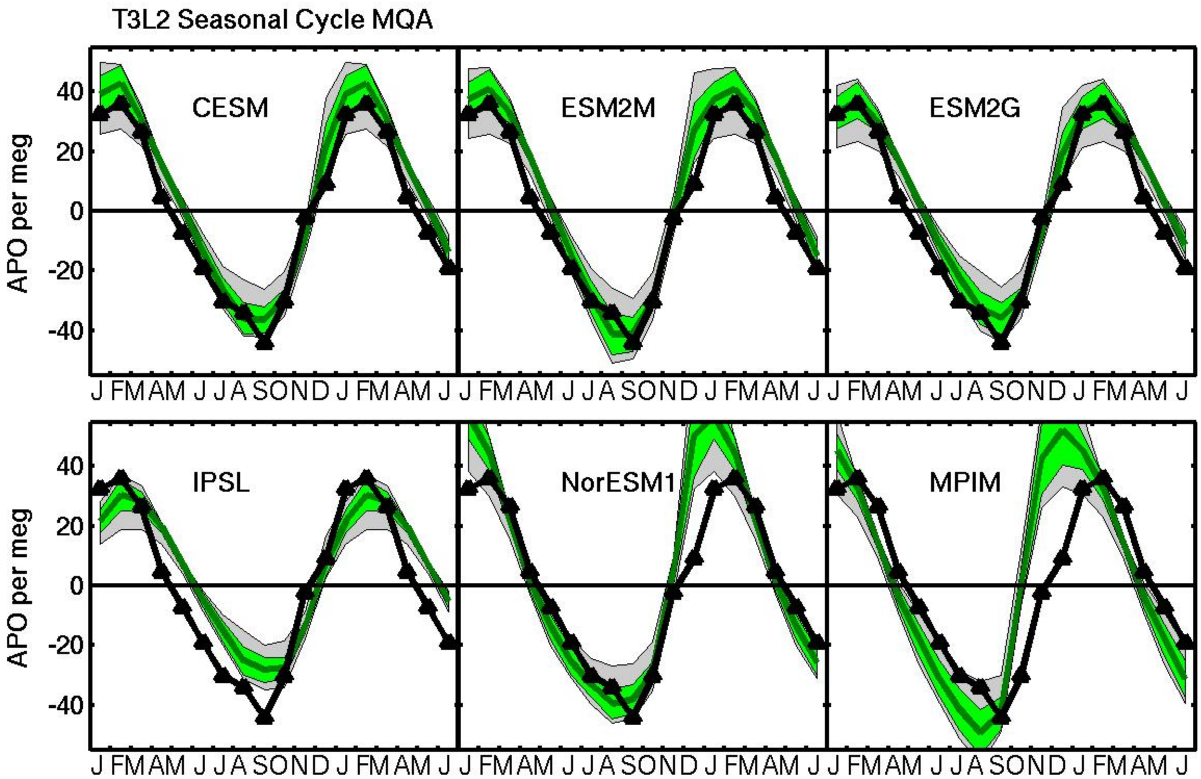


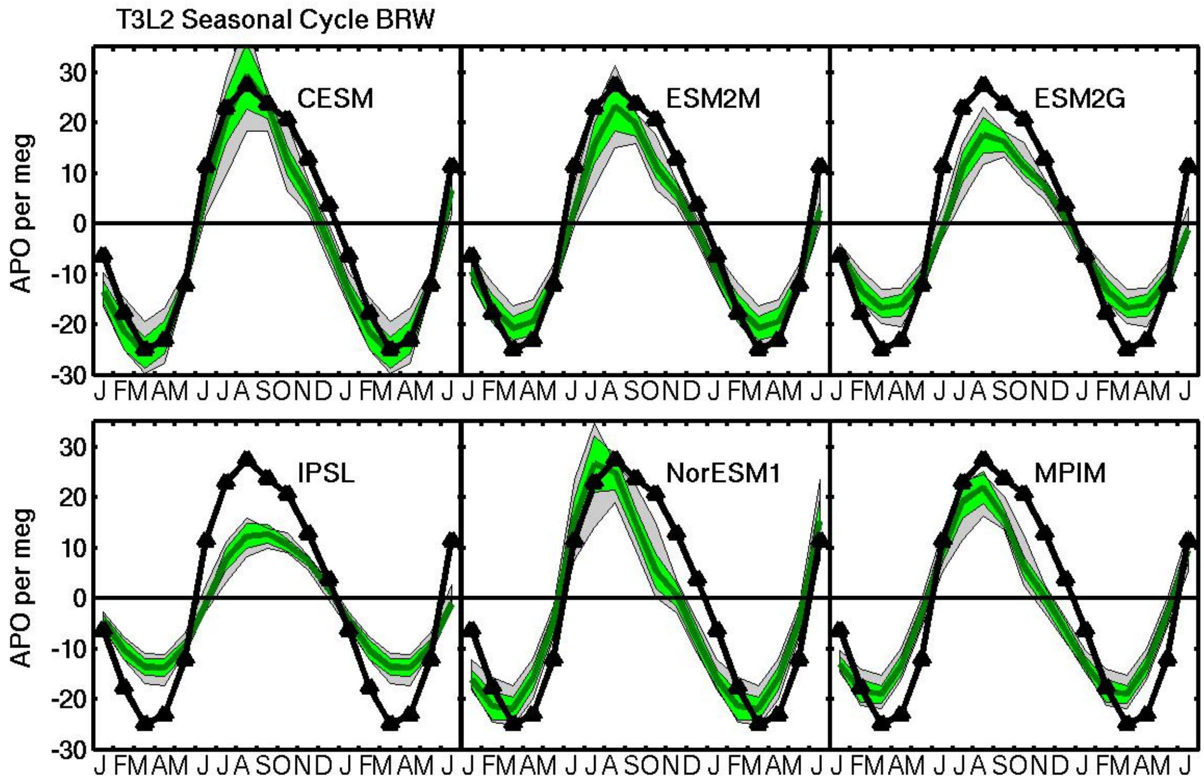
Figure 2. Results of the pulse-response code forced by O_2, N_2 and CO_2 air-sea fluxes from 6 ESM ocean biogeochemistry model components. The dark green line and light green window show the mean and standard deviation, respectively, of the 9 ATMs participating in both T3L2 and APO Transcom. The amplitudes are scaled for each ATM and monitoring site based on the validation exercise described in Section 2.2.2 and illustrated in the Supplementary Material. The gray window shows the full range of responses from all 13 T3L2 ATMs, uncorrected based on the APO Transcom validation exercise. The heavy black line shows the observed APO mean annual cycle. a) Results at South Pole, compared to SIO observations.



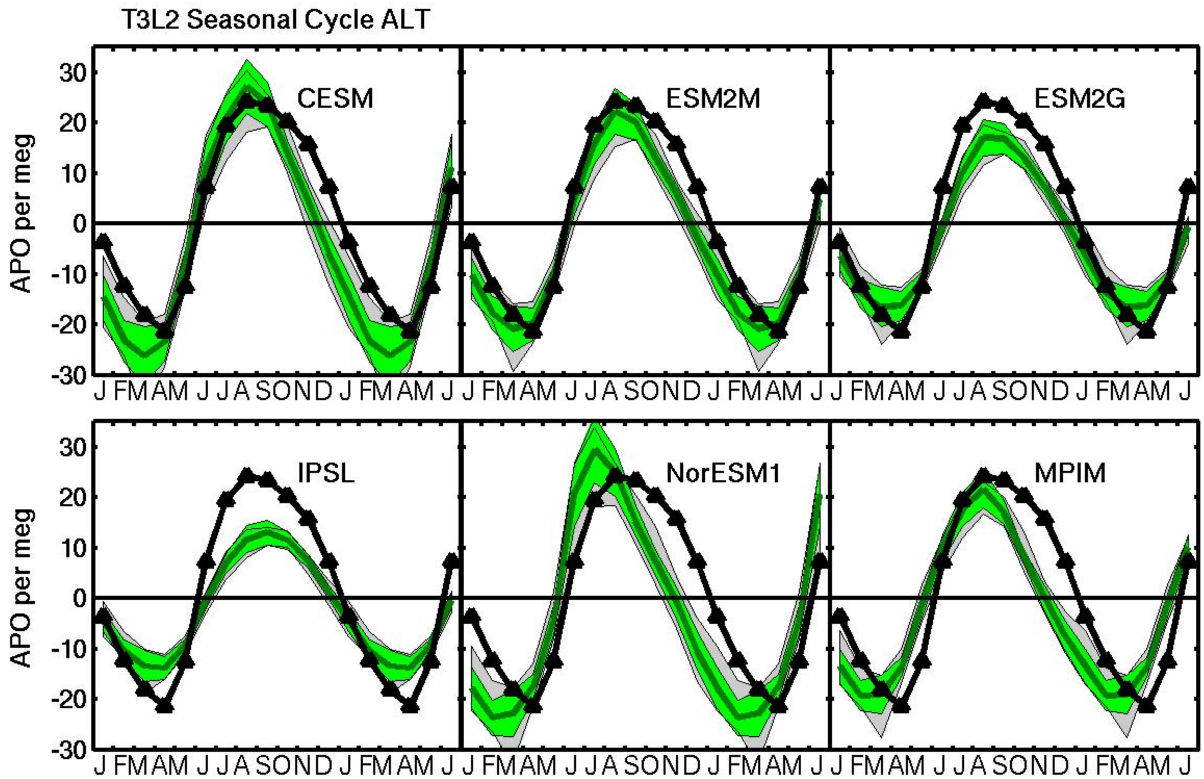
2b) Results at Palmer Station (64.9°S, 64°W), compared to SIO observations.



2c) Results at Macquarie Island (54.5°S, 159°E), compared to PU observations.



2d) Results at Barrow, Alaska (71.3°N, 156.6°W), compared to PU observations.



2e) Results at Alert, Canada (82.5°N, 62.5°W), compared to SIO observations.

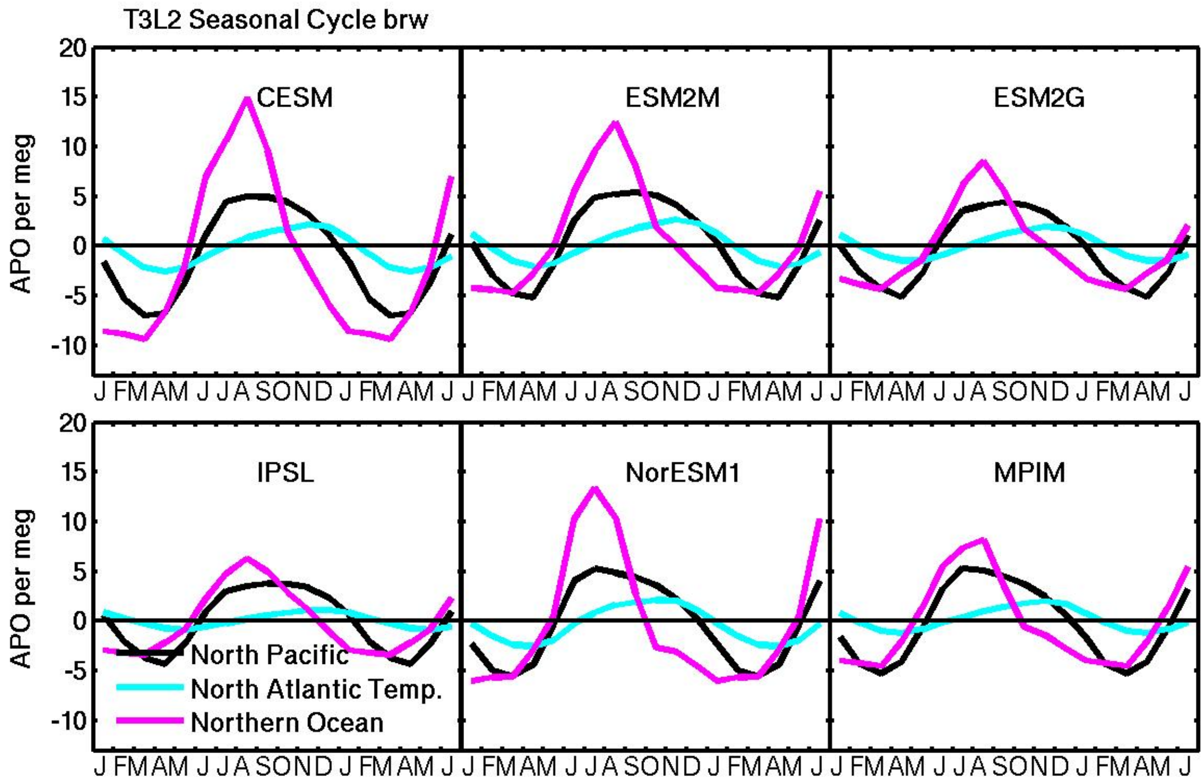


Figure 3. Partitioning the APO cycle at Barrow, Alaska into its main regional contributions, North Pacific (black), Temperate North Atlantic (cyan) and Northern Ocean (magenta), which includes the North Atlantic north of 48N and the Arctic Ocean. All curves reflect the unscaled model mean of the 13 ATMs used in the matrix method.

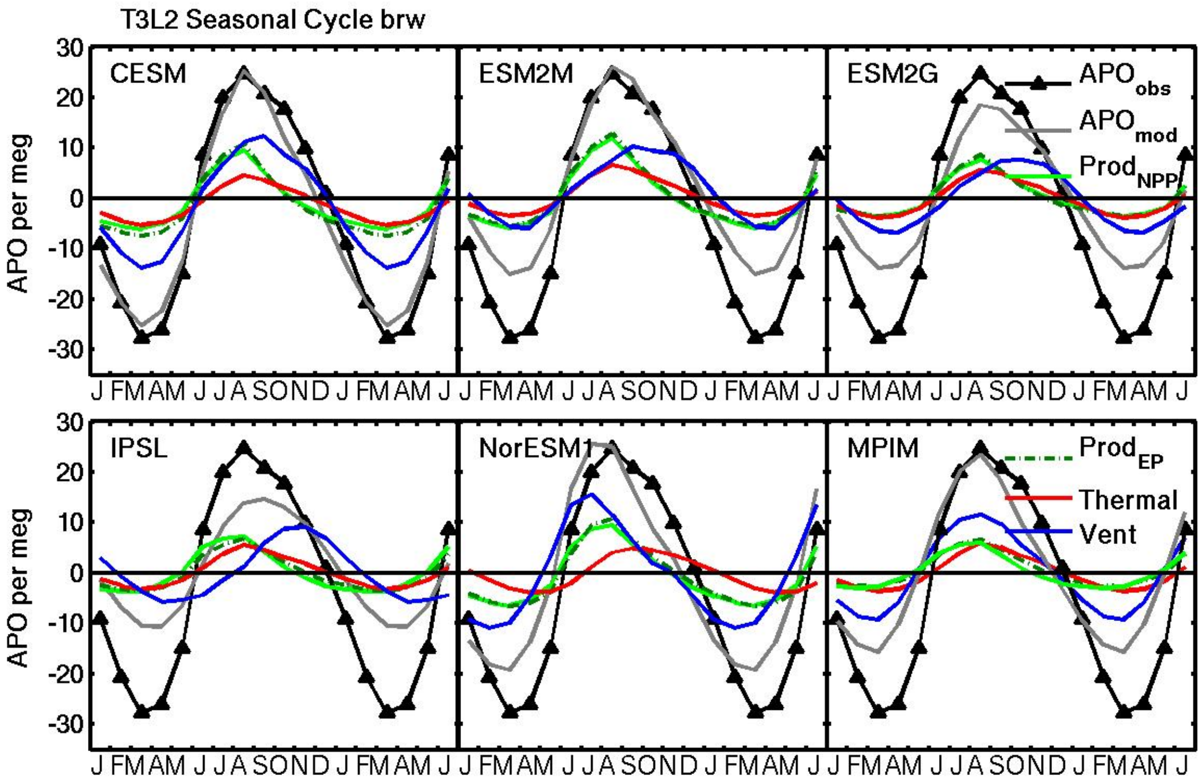


Figure 4. Partitioning of the model mean APO cycle into NCP, thermal and residual ventilation components at Barrow, Alaska. The APO_{NCP} components are estimated alternatively based on ocean model EP at 100m ($Prod_{EP}$ light green, solid curve) and vertically-integrated NPP ($Prod_{NPP}$) scaled by the mean ratio of EP_{100}/NPP (f ratio) between 40-60°N of the given ocean model (dark green, dashed curve). All components were translated into atmospheric signals as described in section 2.2.3. Also shown is APO_{vent} (blue), calculated as a residual of APO – APO_{NCP} – APO_{therm}. With the exception of observed APO, all curves reflect the unscaled mean of the 13 ATMs used in the matrix method.

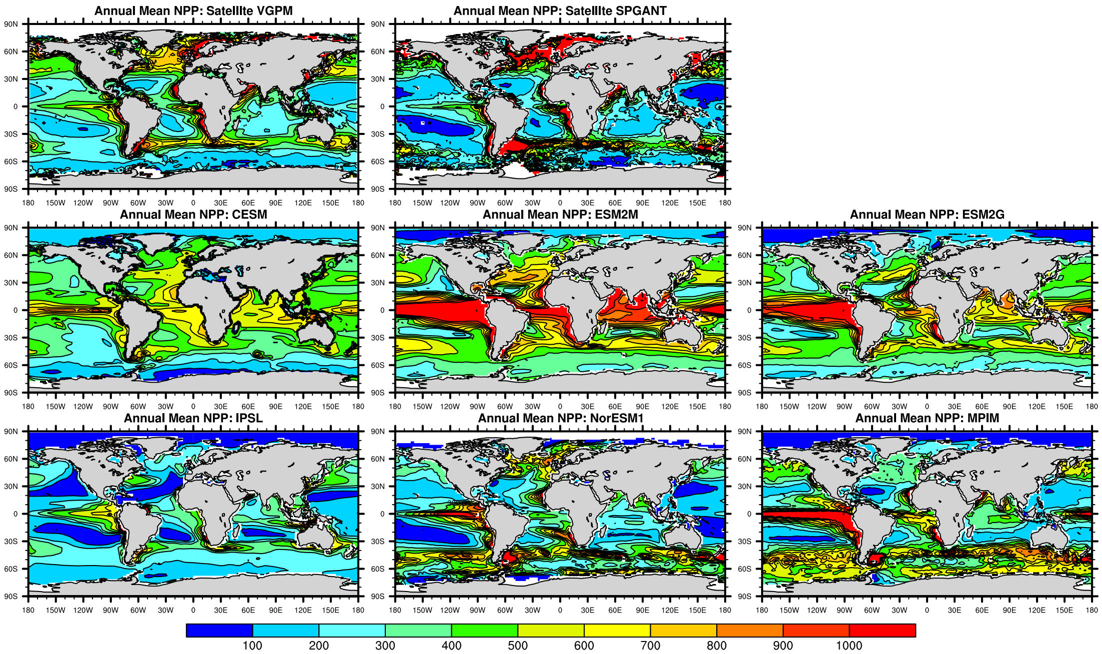


Figure 5. Annual mean NPP (in mg C m^{-2} day^{-1}). Top row: MODIS-Aqua data input to the VGPM NPP model and b) SeaWIFS data input to the SPGANT algorithm as described in *Nevison et al.* [2012]. Rows 2 and 3 show the corresponding NPP fields from 6 ESMs for the mean of 1997-2005.

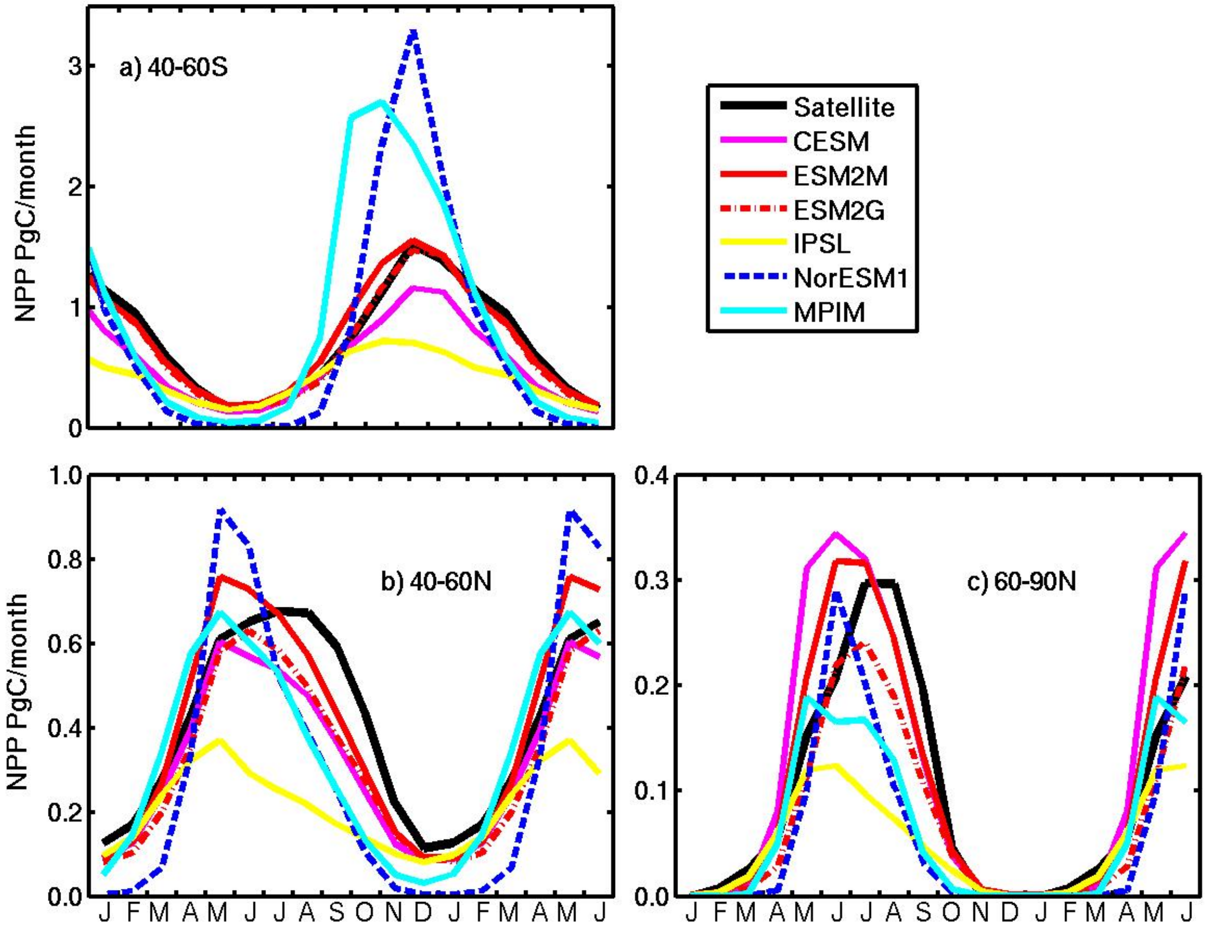


Figure 6. Comparison of the NPP (PgC month) mean annual cycle as simulated by ESMs and satellite-derived observations integrated over: a) 40-60ºS, b) 40-60ºN, c) 60-90ºN. The satellite data are from SPGANT/Laws in panel (a) and VGPM/Dunne in panels (b-c).

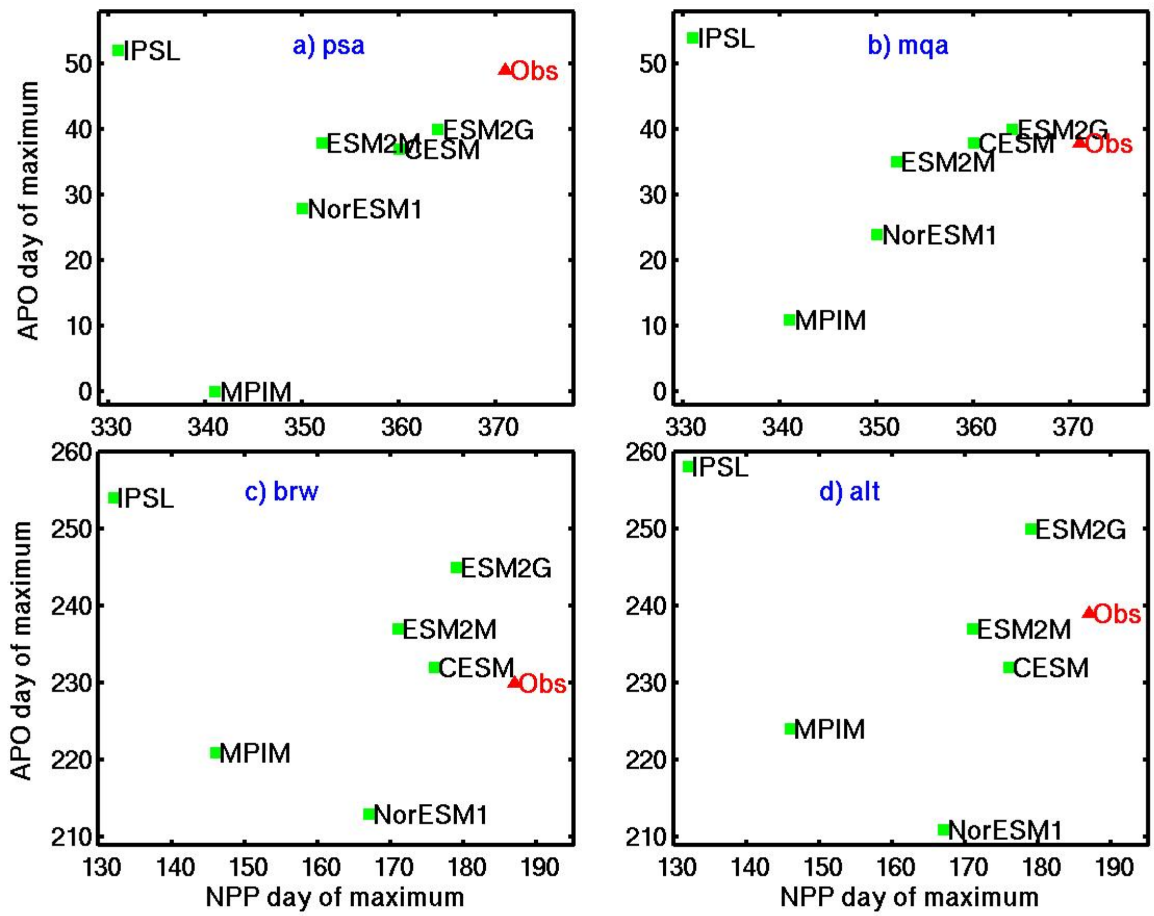


Figure 7. Day of APO maximum plotted against day of NPP maximum. The observed data point is derived from APO data at a) Palmer Station, b) Macquarie, c) Barrow and d) Alert plotted against satellite NPP data integrated over the 40-60 degree latitude band of the appropriate hemisphere.

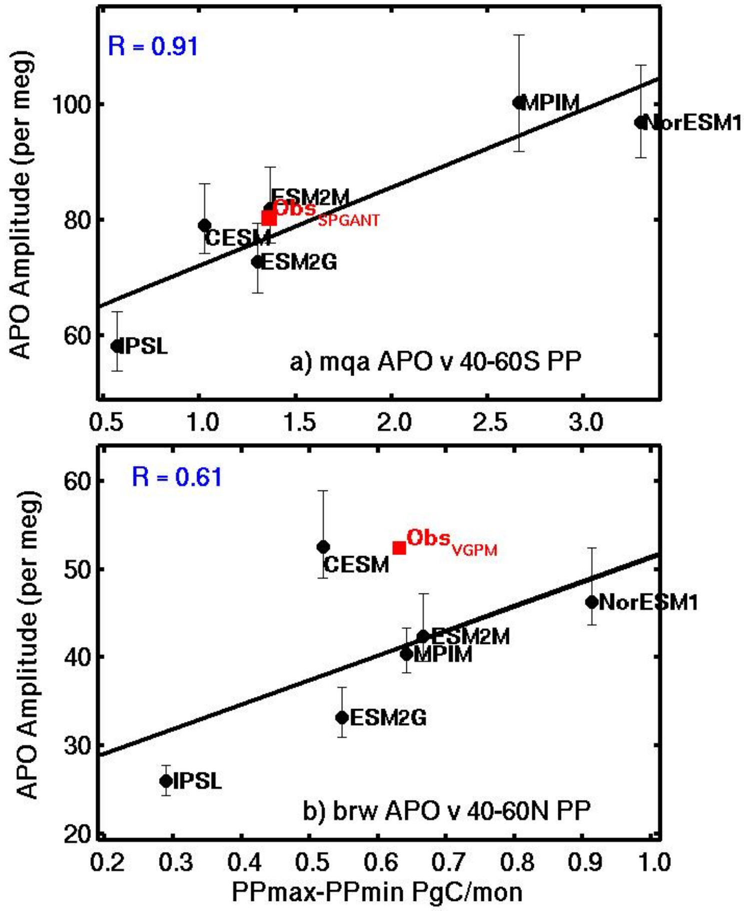


Figure 8a. Seasonal amplitude in APO at Macquarie Island (MQA), located at 54.5S, 159E, as estimated from the air-sea O_2, CO_2 and heat fluxes from 6 ESMs, plotted against the seasonal amplitude of NPP integrated from 40-60S. Error bars represent the ATM uncertainty in model APO as estimated with the matrix method. The "Observed" data points (in red) are based on APO data from the PU network at Macquarie and NPP from the SPGANT satellite ocean color algorithms, as described in the text. The correlation coefficient R refers to regression through ESM points only, b) Same as 8a, but plotting seasonal amplitude in APO at Barrow, Alaska against the seasonal amplitude of NPP integrated from 40-60N. The "Observed" data point is based on APO data from the PU network and the VGPM algorithm with MODIS-Aqua input.

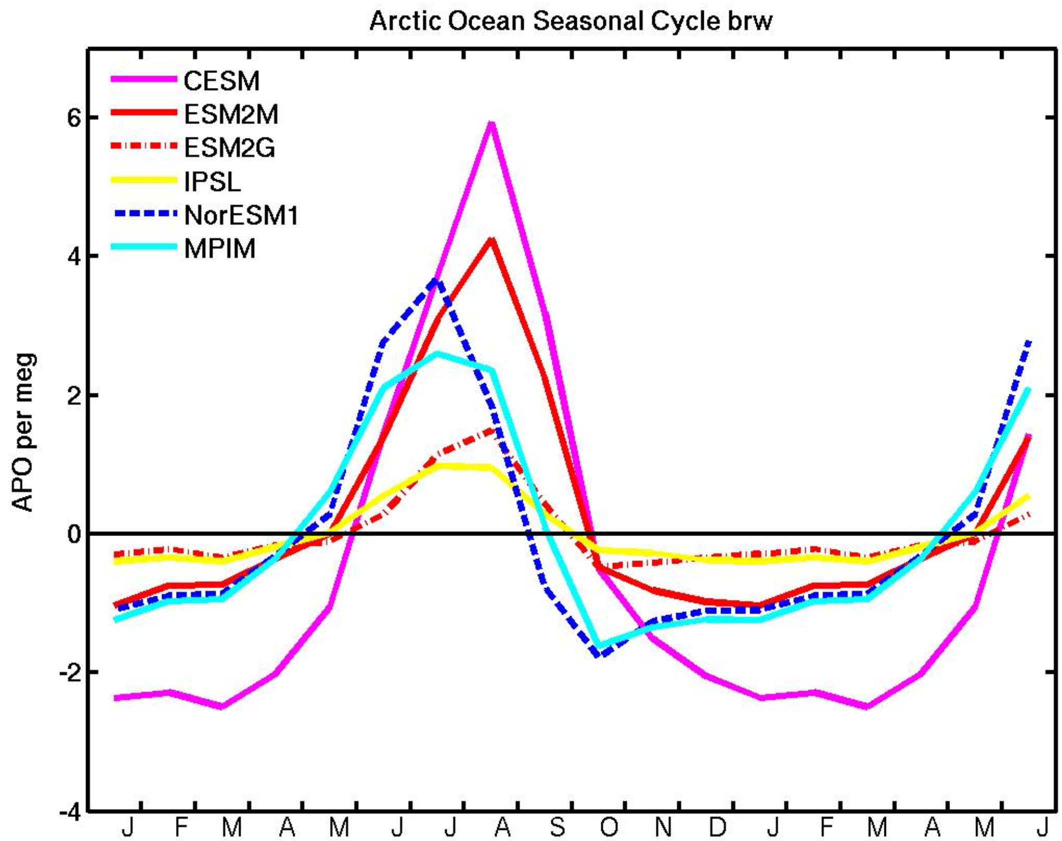


Figure 9. APO cycle at Barrow, Alaska from the Transcom Northern Ocean region, restricted to latitudes north of 65ºN to estimate the contribution of the Arctic Ocean. All curves reflect the unscaled model mean of 13 ATMs used in the matrix method.

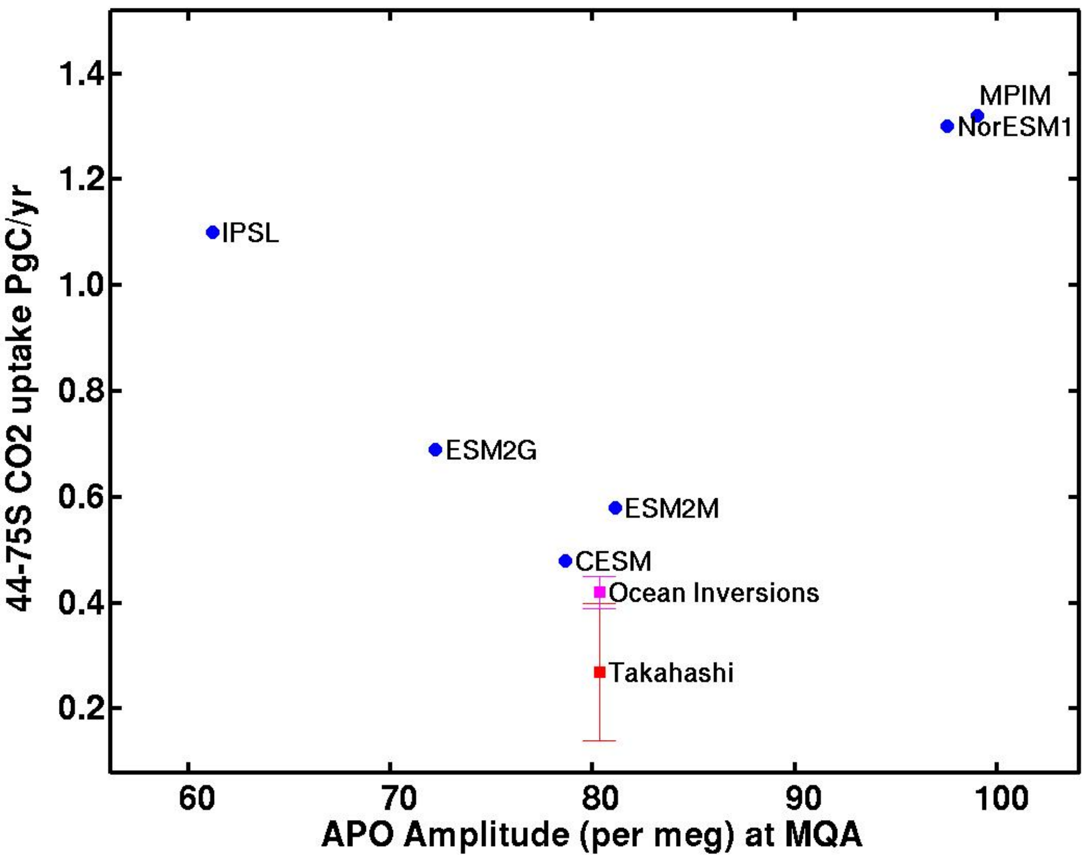


Figure 10. Annual mean CO_2 uptake in the Southern Ocean for 1997-2005 integrated 44-75°S plotted vs. mean APO amplitude at Macquarie over the same period, as predicted by 6 ESMs. Independent estimates of carbon uptake from ocean inversions and observed pCO_2 databases [*Lenton et al.*, 2013], plotted against the observed APO amplitude at Macquarie are shown for reference