Response to Reviewer 1 (Britt Stephens).

We thank Dr. Stephens for his helpful and very detailed comments. His major comments are reprinted here in blue and our responses are given in black. Below, we also provide a list of responses to the minor comments in his annotated PDF file.

This paper applies observations of seasonal cycles in atmospheric oxygen to evaluate a subset of ocean biogeochemistry models participating in the CMIP5 project and uses satellite-based productivity estimates to derive complementary insights. It is a nice demonstration of the applicability of oxygen data to this task and provides useful insights into the behavior of recent models relative to observations and other models. Although productivity estimates from space are highly uncertain, the paper shows that phasing information makes an additional contribution. The use of matrixed response functions from TransCom3 era uniform flux simulations to link the models and observations is not optimal, but I recommend publication with only modest revisions. Major comments are below, while minor suggestions are made inline in the attached pdf.

Major comments

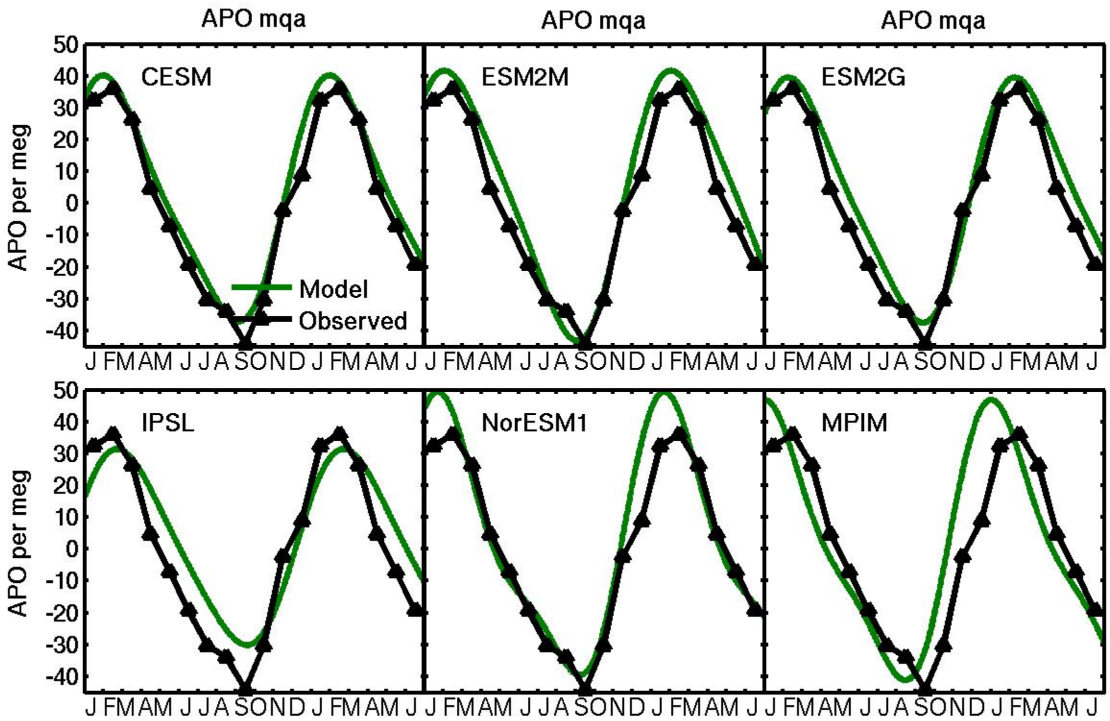
1) The authors use atmospheric transport simulation output from the TransCom 3 Level 2 experiment to translate ocean model fluxes into estimated atmospheric signals. These model runs were conducted about 13 years ago and there have been significant advances in atmospheric transport model resolution and fidelity since then. Furthermore these runs were done using uniform flux distributions and as the authors show, this leads to considerable differences with respect to O2 specific patterns. If I were associated with one of the ocean models that doesn't look very good in this analysis, I would be tempted to insist that the analysis be redone with modern atmospheric transport models and O2 flux patterns. Independent atmospheric transport models have also converged significantly in the past decade, and using the TransCom model spread as an estimate of uncertainty here may undersell the potential for atmospheric O2 data to test ocean models. For example, the standard deviation on northern extratropical land fluxes, which has been linked to differences in vertical mixing, shrunk by over a factor of 2 from the TransCom 3 Level 2 study (Gurney et al., GBC 2004) and the RECCAP study (Peylin et al., BG, 2013), and the RECCAP study allowed different methodologies and observational networks suggesting transport has converged even further. Of course, the right thing to do would be to collaborate with atmospheric transport modeling groups to run O2 flux patterns through modern transport models. Using these old matrixed response functions, which as the authors point out can be run in seconds, seems somewhat to be taking the easy way out. Nonetheless, the approach and results presented here are sufficiently well defended for publication. I would however suggest adding discussion of the dated nature of these simulations and the possibilities of bias and/or overestimated uncertainty. I would also encourage the authors to use more rigorous atmospheric transport simulations in future work.

The matrix method was a deliberate effort to address criticism raised in the literature (e.g., by Naegler et al., 2007, Battle et al., 2006, and indeed Stephens et al., 1998) that ATM uncertainty reduces the confidence one can place in APO as an evaluation metric for ocean model air-sea fluxes. Some of those papers went so far as to suggest that the uncertainty is so large that APO does not provide a useful constraint. Our matrix method provides a means to quantify the ATM uncertainty, although it likely does tend to exaggerate that uncertainty (the use of the best guess green envelopes and broader gray envelopes was an attempt to show that the most likely range of uncertainty is narrower than the full width of the gray envelopes). Peylin et al, 2013 and other RECCAP papers show *a posteriori* inversion results for CO2, so a number of assumptions are needed to cite these papers as evidence that the current generation of ATMs will have converged on

APO relative to the T3L2 models. Further, at least some of the T3L2 are still actively used (e.g., TM3), which makes it a bit awkward to suggest that these ATMs are outdated. In general, we feel some reluctance to undermine our T3L2 matrix approach based on speculative arguments about reduced ATM uncertainty in APO using modern ATMs.

In defense of our matrix method, Transcom3L2 involved a substantial international effort and coordination that, to our knowledge, has not been repeated since. As part of Transcom3 L2, 13 different ATM modeling groups ran simulations with the same surface forcings to generate a large, publicly available database of standard output files, including the pulse-response functions used in our matrix method. The Transcom APO exercise was a spinoff of T3L2 that provides a means for linking and evaluating the T3L2 basis functions to forward simulations of APO with most (9) of the same 13 models. In comparison, the RECCAP effort cited by Reviewer 1 was considerably less standardized and had no obvious connection to APO. It involved "Eleven sets of carbon flux estimates … generated by different inversions systems that vary in their inversions methods, choice of atmospheric data, transport model and prior information." While the matrix method used here can be criticized on a number of levels, in the absence of a new, internationally coordinated effort that is beyond the scope and resources of our present work, the pulse-response functions generated by the Transcom modelers provide the most readily available means to compare uncertainty in modeled APO among a wide range of ATMs.

That said, we have added the following sentences to Section 4.2: "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, potential reducing it substantially from the broad windows that result from our current matrix approach." We also have performed some full forward simulations with GEOS-Chem, a modern-era ATM that has been used extensively in CO2 passive tracer simulations, and obtained results that are generally consistent with our matrix method.



Review Response supplementary figure 1. APO at Macquarie Island computed from forward simulations of the GEOS-Chem model forced by 1994-1997 O_2, N_2 and CO_2 air-sea fluxes from 6 ESM ocean biogeochemistry model components (green curves). Black curves show the observed APO mean annual cycle. The results obtained from these forward simulations with a single ATM are largely consistent with the results obtained from our matrix model method based on the T3L2 pulse response functions. The top row ESMs capture observed APO relatively well, while the bottom row ESMs do not.

2) This study evaluates 6 ocean biogeochemistry models that were part of CMIP5, but there were more participating models and the text does not explain why these 6 were chosen. Is there something special that distinguishes them from others? If this work is intended primarily as a demonstration of a method, then 6 models is sufficient, but this should be explained clearly in the introduction.

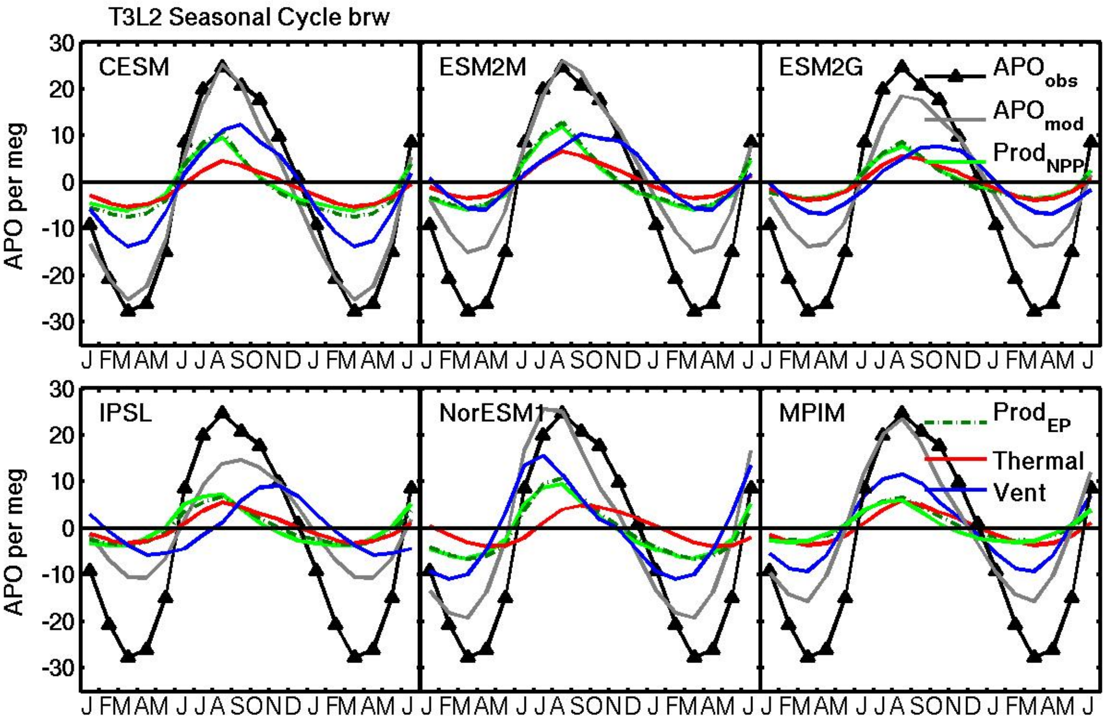
We explain more explicitly in Section 2.1, that, “Many of these (needed CMIP5 output) fields were available through public web interfaces, but some variables, particularly Q, required assistance from the individual modeling groups, which effectively limited the study to 6 models listed above.” We have also stated in the Introduction that, “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.”

3) Equation 2 parses FO2total as the sum of FO2ncp, FO2vent and FO2therm. The authors have confidence in FO2therm and 2 methods for estimating FO2ncp. FO2vent is then estimated as a

residual and the authors conclude that in many places it is unreasonable so they do not use it. However, if FO2vent calculated as a residual is wrong, then one of the other terms must also be wrong. Unless the authors can explain how FO2vent as a residual could be wrong while FO2ncp is right, then I don't think FO2ncp should be used individually either. Rather, a combined FO2bio should be calculated as a residual and used.

Partitioning APO_{bio} into APO_{NCP} and APO_{vent} components was an important goal of this paper, because isolating APO_{NCP} is the most straightforward way to compare APO directly to satellite ocean color data (see discussion in Nevison et al., 2012a). Rather than showing only APO_{bio}, we think it is more useful to at least attempt the partitioning, and then discuss why it may be falling short in some regions (like the Southern Ocean).

To specifically address the reviewer's comment, we now include the APO_{vent} term in Figure 3 (now Figure 4) (at Barrow, AK), while including caveats that,
"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 NorESM1 curve is clearly unreasonable in phasing)."

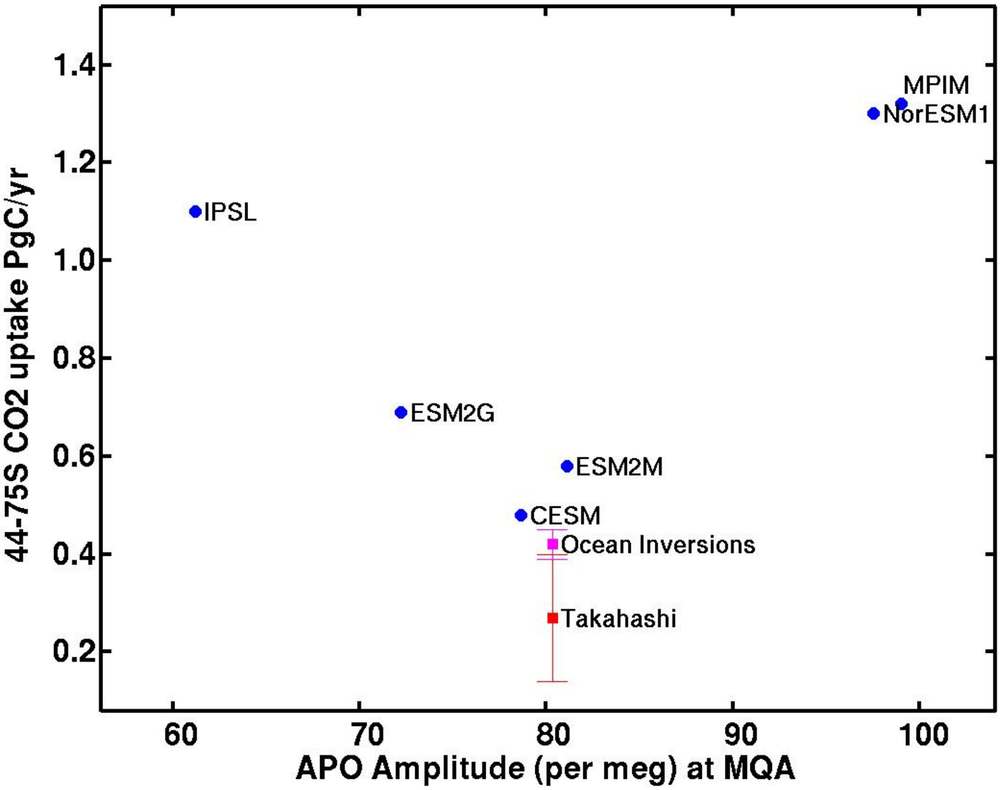


New Figure 4, partitioning APOncp, APOtherm **and APOvent** at Barrow.

At the end of Methodology Section 2.2.3 we also have added text to clarify the rationale for considering APO_{NCP} in the Southern Ocean while avoiding APO_{vent}, "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."

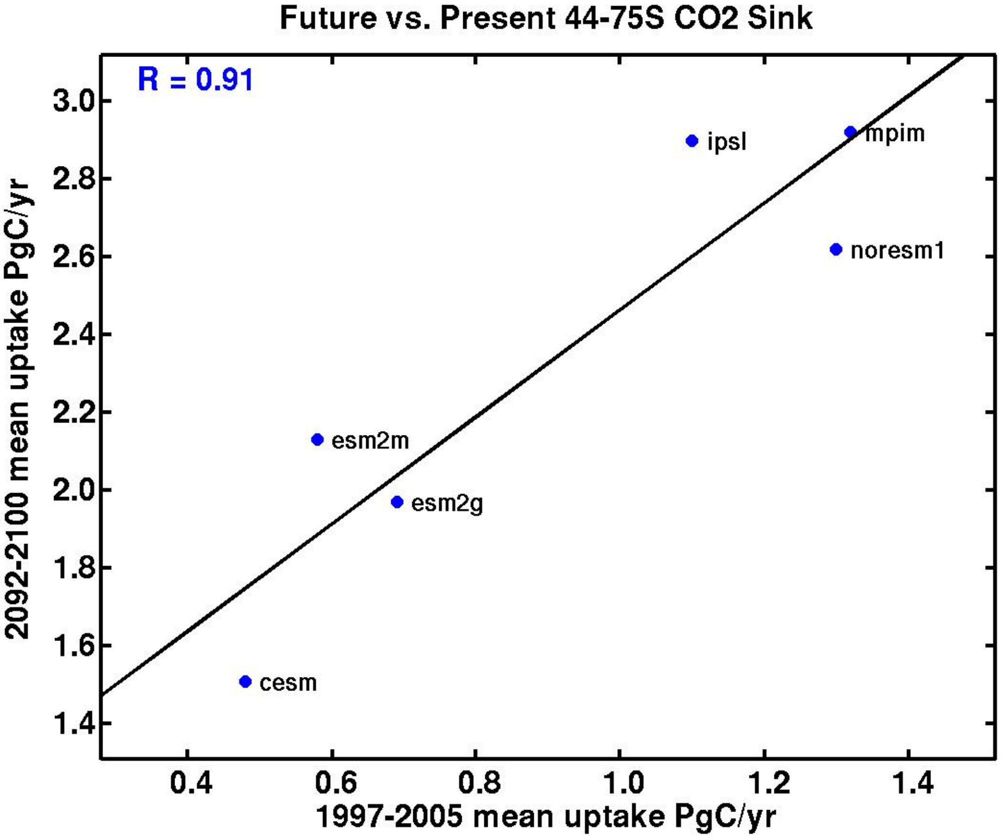
4) Some discussion of the relevance of the model-assessment results discussed here for assigning confidence to future carbon-climate projections by these models would be valuable. Are the poor-performing models at all distinct in their projections of future CO2 uptake by the ocean? Does this method have promise as a tool for evaluating future climate projections?

We have included a new Figure 10 that addresses this question, at least with respect to present day ESM prediction of CO2 uptake in the Southern Ocean.



The **new Figure 10** shows annual mean CO_2 uptake in the Southern Ocean for 1997-2005 integrated from 44-75°S and plotted vs. mean APO amplitude at Macquarie over the same period. We discuss in Section 4.2 how the ESMs that reproduce APO the best in the Southern Ocean tend to predict a smaller present day net carbon uptake between 44-75° than those (IPSL, MPIM, NorESM1) that perform more poorly on APO. As shown in Figure 9, the top performing models on APO are also in better agreement with independent estimates of carbon uptake from ocean inversions and observed pCO_2 databases [*Lenton et al.*, 2013].

Reviewer Stephens also asks about future CO2 uptake. Since our current manuscript focuses on the historical (1850-2005) CMIP5 simulations, this question is probably beyond the scope of the present work. However, we note here that our further work with the RCP8.5 future scenario, based on mean results from 2092-2100 for the same 6 ESMs, suggests that present day and future CO2 uptake are well correlated. This suggests that the models that perform poorly on CO2 uptake in the present day may tend to overestimate future Southern Ocean CO2 uptake.



Review response supplementary figure 2, showing annual mean CO_2 uptake in the Southern Ocean for 1997-2005 integrated from 44-75°S compared to annual mean CO_2 uptake from 2092-2100 under the RCP8.5 forcing scenario.

5) If the only information coming from satellite ocean color is phasing, would it not be simpler to just use satellite NPP, which presumably has very similar phasing? Some discussion of the value of satellite NCP estimates in the context of phase information only (if there is one) would be useful.

We have effectively done this by using satellite NPP rather than NCP/EP in Figures 5-8. However, since NCP/EP is in principle more closely related to APO_{NCP}, we think it is useful to consider both quantities (as in Figure 4). Discussing the relationship between satellite NCP and EP also provides a background for one of the points in our Conclusion, “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]. Extending model-derived insights to satellite products may be more challenging and will likely 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."

Response to minor comments annotated in the text. We have followed all the reviewer's suggestions, unless specifically noted. Since many of the suggestions are minor wording changes, we only explicitly respond to the comments that required substantial changes:

p.6 comment 1: We have added, "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 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.
".

p.7 comments 1-3 were addressed by rewriting paragraph 3 of Section 2.1 as follows:
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.
*other CMIP5 intercomparisons (e.g., Anav et al., 2013) do not provide explicit references or meteorological drivers for the historical simulation. We have followed this precedent of describing the historical simulation as one that is driven by a standard prescribed atmospheric CO2 concentration from 1850-2005. As for the meteorological drivers, these are model specific and thus described in the individual references for each model.
** while these web interfaces exist, in practice we received most of the output directly from the modeling groups, and therefore have included representatives from each group as coauthors (except in one case where the modeler preferred to be included in the acknowledgements). This is the main reason the study was limited to six models. We also felt that those models provided a sufficient range of results to illustrate the use of APO as an evaluation metric.

p.8 comment 2 – we have cited Gurney et al., 2003, in which their Fig. 1 provides a map of the 11 ocean regions from Transcom3.
p.8 comment 3 – We have clarified that Transcom 3 uses an annually repeating cycle of meteorology. "…using an annually repeating cycle of meteorology that was model specific for each ATM" Table 1 in the cited Gurney et al. 2003 lists the meteorological drivers for each model.
p.10 comment 1. Clarified that we used, "station output from the forward ATM simulations of the APO Transcom Experiment."

p.10 comment 2. Inserted, "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 …"

p.12 comment 1. Since Section 2.2 is a methodology section, we have tried to avoid presenting results here, but do now cite Table 1, which gives the range of *ef* values.

p.12 comment 2. We have decided to show the APOvent in Figure 4, since omitting it seems to be causing more consternation than simply showing it would. We have also modified the text to explain the rationale for rejecting APOvent in the Southern Hemisphere while still considering the other component terms (again trying to defer the presentation and discussion of results to later sections), "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."

p.13 comments 1 and 3, replaced with "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 ..."

p.13 comment 4, we have added,
"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."

p.14 comment 1. This sentence is now included "Details of the station locations and time spans of data used to calculate the mean seasonal cycle are listed in Table S2. 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)."

p.14 comment 6. We include the following text, "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. We reiterate that 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."

p.15 comment 1. The reviewer is probably right and we have deleted this sentence. The proposed alternative method for quantifying uncertainty involves an analysis of IAV, which, as stated above, is beyond the scope of the current study.

p.17 comment 2 (see also response to p.12 c1) We have replaced the highlighted text with, "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 3 should be interpreted with caution (e.g., the NorESM1 curve is clearly unreasonable in phasing)."

p.18 comment 1. We have added this introductory statement, "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."

p.20 comment 2. We have added, "The inference from the APO component analysis in Figure 3 that the GFDL models may have weak ventilation in the North Atlantic …"

p.22 comments 1-5. We have rewritten as, "we currently are not able to distinguish which of the underlying air-sea O_2 flux fields is the most accurate, 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 7a). Reducing ATM uncertainty is a challenging problem 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, potential reducing it substantially from the broad windows that result from our current matrix approach."**

p.22 comment 6. We have added, "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."

p.24 comment 1. We have added, "The first of these, ATM uncertainty, is large, as quantified using our Transcom3-based matrix method, but probably also has been overstated in previous analyses [e.g., *Naegler et al.*, 2007]. ATM uncertainty also may be reduced substantially in future work with modern ATMs and O_2-specific flux patterns."

p.33 comments 3 and 6. "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 Transcom APO validation exercise. The heavy black line shows the observed APO mean annual cycle. a) Results at South Pole, compared to SIO observations."

p.40 comment 2. We have removed the illegible labels from the tops of each panel in Figure 4.

p.41 comment 1. We have added this to the Fig 5 caption, "The satellite data are from SPGANT/Laws in panel (a) and VGPM/Dunne in panels (b-c)."

p.42 comment 3. We have moved the labels to the top of the panels and enlarged the font.

p.45 comment 1. We have deleted the sea ice figure and replaced it with a new Figure 9 that addresses major comment 5 – relating ESM performance on APO to carbon uptake in the Southern Ocean.

Supplementary Material.

Page 2, comment 1. We have provided more information about the APO Transcom forward simulations (FS):

"In contrast to the matrix-based PRC simulations, which used uniform regional distributions of O_2 and N_2, the archived APO Transcom forward simulations were forced by fine-scale (0.5 x 0.5 degree) monthly mean air-sea flux distributions (interpolated by APO Transcom from the original 1.125 degree resolution of *Garcia and Keeling* [2001]). The simulations were run by each participating model group with the fluxes turned on for the first year and turned off for the last two years. The resulting ATM atmospheric O_2 and N_2 fields in ppm were sampled in each of the 36 months of the simulations at 253 monitoring sites. The steady-state response, i.e., the mean seasonal cycle, was computed by summing all Januaries, Februaries, etc., for the three years. Conceptually, this calculation assumes that the ATM behaves linearly and that the steady-state response can be represented as the sum of the response to the fluxes from the present year, the past year, and two years previously, which correspond to the first, second, and third years of the simulations, respectively.

In using the archived APO Transcom results, it was necessary to account for several irregularities. First, the JMA O_2 and N_2 results were multiplied by 10^6 to convert to ppm units. Second, TM3 ran all 36 months with pulses on, so instead of summing all 3 sets of Januaries, Februaries, etc., the mean annual cycle was calculated based on the third year of the simulation alone. Finally, GISS UCI in principle was a 10th model that participated in both T3L2 and APO Transcom, but in practice it could not be used because only the first (pulse-on) year of GISS UCI output was submitted to APO Transcom."

Page 2, comment 2. We speculate as to why the sigma ratios might be < 1 at the bottom of Table S2: "At most extratropical stations, the σ_{prc}/σ_{fs} ratios are < 1, suggesting that the Pulse Response Code tends to underestimate the true APO amplitude from the forward simulations. This may be due to the uniform flux distributions assumed across Transcom regions, which could smooth out hotspots for O_2 air-sea flux that may lead to more intense peaks in true APO."

Page 2, comment 5. We have provided correlation coefficients as R^2.

Page 3, comment 1. Columns added for time period used for 5 stations in Fig. 1

Page 3, comment 3. We have added the missing 3 stations (RYO, CGO5500m, and MLO) to Table S2.
Page 3, comment 4. We have provided correlation coefficients as R^2.
Page 5, comments 1,6. We have added the missing 3 stations (RYO, CGO5500m, and MLO) to the Taylor diagrams and provided the Taylor, 2001 reference