Response to Reviewer 2.

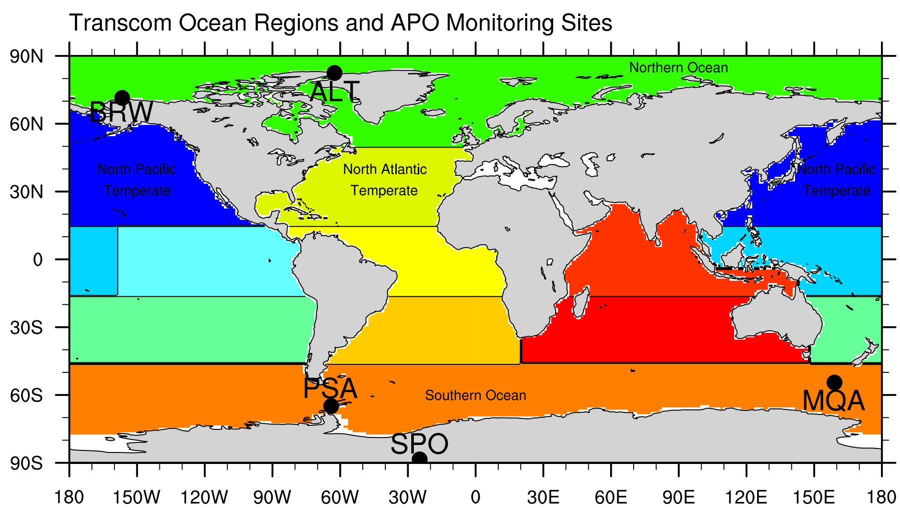
We thank Reviewer 2 for his/her helpful comments, which are reprinted here in blue. Please see our responses in black.

The authors present a comprehensive evaluation of the ocean biogeochemical com- ponents of 6 CMIP5 models against observed APO and Satellite estimates of phyto- plankton productivity. The goal here is to offer the APO datasets, in particular, as a new constraint on the models. The authors use a transport matrix method so as to speed the process of atmospheric transport substantially. They compare this method to a direct method and only consider regions where this works well. Atmospheric trans- port uncertainty is smaller than variance across the ocean biogeochemical models for the high latitude sites. This is important, since the utility of APO has generally been questioned by the fact that one must do this transport calculation. The authors could point this out more clearly, i.e. in conclusions. On the whole, this is a nice analysis that should be published after minor revisions.

Major comments: 1. The transport matrix is a good step, and I support its use for this paper. Going forward, the authors might consider developing such a matrix approach based on regions different from the square boxes of TRANSCOM that do not capture the biogeography of the ocean well. Fay and McKinley (2014) offer global biomes that would be preferable. For this paper, the authors need to clarify if the aggregation across these square biomes could impact their results and the model-to-model differences that are found. Specifically, if models don't have their major biogeochemical gradients across the TRANSCOM region boundaries, could this influence these comparisons? I also ask that TRANSCOM region boundaries be included in at least one panel in Figure 4. Fay, A. R. & McKinley, G. A. Global open-ocean biomes: mean and temporal variability. Earth Syst. Sci. Data 6, 273–284 (2014).

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, Stephens et al., 1998) that ATM uncertainty reduces the confidence 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. The matrix method provides a means to quantify the ATM uncertainty. Somewhat surprisingly, our first reviewer suggested that ATM uncertainty is no longer as important a problem and therefore it would be better to use full forward simulations than the matrix method. While we concede that he may be right, we are also concerned that he may be dismissing too casually the lingering issues with ATM uncertainty, especially since he does not offer direct proof that ATM uncertainty is no longer a major problem for APO analyses. Please see our response to Reviewer 1 for further discussion.

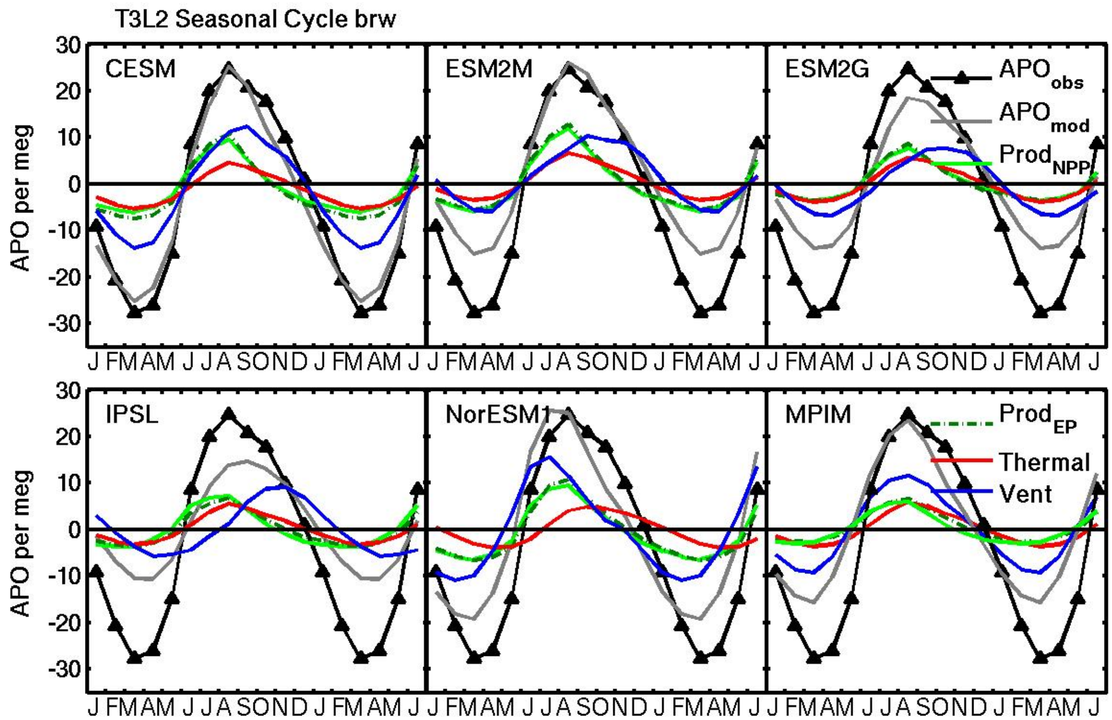
We agree with Reviewer 2 that the latitude-based boundaries of Transcom3, which we now show in our new Figure 1, are not ideal for capturing the main biogeochemical boundaries. The biomes defined in Fay and McKinley, 2014 would likely be an improvement, and the partitioning of the Southern Ocean into 3 different regions based on biogeochemical function, could provide insight into the contribution of these different regions to variability in APO. While it is beyond the scope and resources of the present study to rerun the T3L2 basis functions to create new biome-oriented basis functions, we now discuss the advantages of this strategy in the following text added to the Discussion, "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."



New Figure 1

2. It is unfortunate that the Ventilation and NCP signals cannot be distinguished; and at the same time the NCP estimates from satellite are so uncertain that we have a reasonably loose constraint here. Showing the APOvent estimated as a residual would be helpful in Figure 3 to add to the text discussion and to better highlight this issue.

We now include the APO_{vent} term in 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 (formerly 3), 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."

3. The conclusions state that the major issues are ATM uncertainty and uncertainty in EP100. The paper suggests to me that the ventilation separation is also quite important, and that the ATM transport is a smaller issue at the high latitudes where this paper focuses. The ATM transport issue at lower latitudes may be more an issue of the TRANSCOM region definitions and how to turn a forward model into a matrix transport approach – but this is really more a technical issue with respect to the challenge of running atmospheric models than about uncertainty in ATM transport. Overall in the conclusions, the authors need to clarify better the many issues that they reveal with their analysis so as to leave the reader with a clearer picture of the value of APO in ESM evaluation, and the remaining challenges to increasing its utility. This discussion might be well-served by a clear separation between Northern high latitudes, mid/low latitudes, and Southern high latitudes.

We have revised the Conclusions as follows to address these points:

"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."

Response to minor comments annotated in the text.

p.8488: We have replaced with, "The exported carbon subsequently is respired in the subsurface ocean, leading to O_2 depletion at depth. O_2 is replenished by…".

p.8488 comment 2: We have expanded to, "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... "
p.8489 : We have cited, "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]." (Both citations are already in the References.)

p. 8492. We have added, "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."

p. 8496. We now show the APOvent term in Figure 4 (formerly 3) and have replaced the highlighted text with, "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."

p. 8497, In response to this and another query from Reviewer 1, 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." We have also clarified that NPP was downloaded from http://science.oregonstate.edu/ocean.productivity.

p. 8501, APOvent is now shown in Figure 3.

p. 8504, Have replaced this sentence with, "The inference from the APO component analysis in Figure 3 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."

p. 8527 Figure 7 Y-labels are both now "Amplitude per meg".

p. 8517 we have added a new Figure 1 showing both the Transcom regions and the locations of APO stations featured in Figure 2 (see above).