We thank the two reviewers for their thorough and thoughtful comments. Their input have helped us to strengthen the manuscript in several areas.

We have responded to the two reviewers main comments by rewriting the introduction and discussion. We compared the mismatches in Chl and NCP with a number of other properties and found that mixed layer depth did the best job in explaining the errors in the models. Mixed layer averaged PAR is also a good indicator but follows MLD closely. We added two figures, a MLD climatology, and a more detailed view of Chl at 60°S.

Anonymous Referee #1

This paper compares two relatively coarse resolution models to satellite chlorophyll data and in situ biological O2 air-sea flux estimates. The bulk of the paper is a description of differences between the model and data with a short discussion of multiple potential reasons for data/model mismatch. The paper would be of interest to a wider readership with the addition of model analysis to more conclusively pinpoint the reasons for mismatch or at least a more thorough discussion of the implications of their findings.
This was addressed by rewriting discussions to stronger emphasize mechanisms and by adding a section about MLD.

The purpose of this paper and its intended audience needs to be clarified. Most of the paper involves a detailed description of the model/data comparison shown in a series of 8 figures. However, there is only speculation about the reasons for model/data misfit. The paper would be much more interesting to modelers if it not only pointed out model discrepancies but also contained some analysis that demonstrated the causes. If the intended audience is mainly observational scientists, it would be more interesting if the paper contained greater discussion of the implications of their results. For example, the last paragraph of the abstract summarizes some interesting conclusions, but these are not actually discussed in the body of the paper that I could find.
We addressed this by rewriting the introduction. and by adding a better discussion about mechanisms. We removed the final paragraph in the abstract because we realized that this would require much more work to be addressed thoroughly

The introduction is overly long and poorly focused. Although much of the paper is about model/data chlorophyll comparisons, chlorophyll is barely mentioned in the introduction whereas the bioflux is discussed at length. It's unclear how several sections are relevant to the rest of the paper, such as paragraph 4 on high-resolution eddy resolving models. Also, the section could be significantly shortened just by tightening up the language.
We addressed this by rewriting the introduction to clarify the purpose of the paper.

Some other tests of the feasibility of aggregating data into a climatological year are warranted. Is the authors' conclusion that this is acceptable for their analysis sensitive to the choice of dates examined (15 Nov & 15 Dec vs. other choices)? How different is the timing of the onset of the spring bloom between different years and does the "blurring" of that onset in a climatology affect the comparisons in the paper? Can anything be said about the errors in creating a climatology by examining the satellite chlorophyll data rather than only the model results? Similarly, can the authors demonstrate that the regions chosen are reasonably zonally

homogeneous? Would it be better to mask some of the data near the coast from the zonal averages, since these are influenced by “processes outside the models’ domain”?
Some of the comparisons made are vague. For example, “the magnitude of peak Chl concentrations is simulated rather well.” Could this be quantified? The color scale is quite coarse in the upper range of Chl, so it’s not clear from the figures that red in both the data and model really represents a good fit.
Our main reason for comparing model and satellite Chlorophyll this way is to allow for comparisons with NCP, where sparsity in observations is a major challenge. We added a figure to quantify the magnitudes of respective Chl range.

Most of the “Conclusions” section does not discuss conclusions from the broader paper, but functions more like an appendix to support the choice of comparing model / data bioflux rather than D(O2/Ar). This should be separated into an appendix and the conclusions section strengthened to discuss implications of the paper’s findings.
We improved the conclusions in correspondence with the rewritten discussion.

O2 supersaturation relative to Ar is defined with different notation from other papers here. It would be better to stick with D(O2/Ar) or DO2/Ar rather than DO2Ar (where D represents Delta here).
We changed the text accordingly.

Paragraph 9 of Introduction. The mixed layer “biological” O2 supersaturation . . .
We changed the text accordingly.

Paragraph 10 of Introduction. Cite some of the “In some studies, . . .”
We rewrote this paragraph.

Section 2.1. Given that the original grid is much finer than the aggregated model grid, it’s confusing to me why there are more holes at the coarser resolution. Do the authors require a certain percentage of good pixels within the aggregated model grid to not disqualify the observations?
There are fewer holes at the coarser resolution, we changed our phrasing to make this clearer.

Figures 7-10, panel d. It would be better to split the map at a different longitude so that the New Zealand data does not appear on both sides of the map, maybe 120oW rather than the dateline.
We believe that the current setup simplifies comparisons with the other maps and that a different projection would be confusing. We are, however, happy to change if the editor prefers the reviewer’s suggestion.

Anonymous Referee #2

This paper presents results of two coarse-resolution ocean global resolution general circulation models (OGCMs) for 4 sectors of the Southern Ocean in comparison to climatologies of satellite-derived chlorophyll from MODIS and a climatology developed from ΔO2/Ar and biological O2 flux observations collected on multiple cruises between 1999 and 2009. The introduction and motivation of the paper are compelling; problems with large scale optimization of OGCMs and how this restricts estimation of smaller scale mixing and seasonal-scale biological processes are presented. As dissolved O2 cycling is affected by both mixing and biological processes at sub-seasonal scales, the DO2/Ar tracer could provide some means of diagnosing how coarser resolution models perform. On the other hand, assumptions of the equivalence of bioflux and NCP are known to be in error in regions of substantial mixing with subsurface waters, and perhaps models could be used to diagnose the error introduced by these assumptions.

Unfortunately, the rest of the paper does not seem to present as clear of a message. On the whole, the models seem to replicate only the ranges of chlorophyll and bioflux for each latitude, but they seem unable to replicate the timing (in either chlorophyll or bioflux) in any sector. The authors suggest that the overlap in range is a signal that the processes responsible for determining NCP are well constrained in the model, but looking at Figures 3-10 I would have to disagree. Isn't a right answer at the wrong time still a wrong answer?
This is of course fundamentally true. However, models are used for different purposes and has to be evaluated in relation to these usages. Seasonal dynamics might not be that important if the aim is to predict basin-wide decadal trends. Likewise, a model that resolves the gulf stream well but locates it's position somewhat too far south can still be of great use. The fact that the models provide a reasonable range give us an important insight in where the models are in error since it suggests that ecosystem processes in the region are constrained rather well. The timing issue is important for many current scientific questions such as the dynamics around spring bloom initiations , the seasonality of CO2 uptake, and how marine ecosystems might be affected by climate change. The errors in timing shown in this manuscript could be connected to physical processes such as mixed layer dynamics and vertical transport of nutrients, or possibly how phytoplankton uptake kinetics are parameterized. We believe that this manuscript provide a strong argument for the lack of specific physical processes in these types of global models, which could be one explanation, and that our results can be a first step. It is of course necessary to further evaluate the models and better assess the exact sources of errors. Such analysis is, however, a significant undertaking and unfortunately beyond the scope this manuscript.

Much of the paper is spent pointing out the many inconsistencies between the models and observations, but not much time is spent discussing the overall trends and what they might mean in terms of model performance. Very sweeping statements regarding construction of the ecosystem in each model or differences in mixing parameterizations are offered as potential explanation for model underperformance, but the discussion ends there. I'm left wondering what we learn from this exercise. In the end, I don't feel the stated objective of the paper (page 9635, lines 23-26) is met. A discussion for what the results "tell us about net community production and the summertime exchange between the mixed layer and the mesopelagic" seems to be lacking.
We have reworded the introduction and discussion to address the reviewers comments.

I feel the approach used here, dissolved gas modeling and the use of tracer-based constraint, is important work, but I just found the paper left the reader wanting for some more mechanistic insight. Instead, I'm left with the sense that the coarse-resolution models are capable of replicating only the most basic of patterns, and there is no real clear indication as to what might fix this. If the authors could develop that side of the story a bit more robustly, I think it would be a much stronger paper, and one that would be well-cited in future studies.

A few more specific comments are offered below:

Abstract, 9630, lines 21-25: these statements are interesting, but they do not seem to be actually discussed in the manuscript.
We have removed these statements since we discovered that a thorough discussion would be to extensive.

9633, lines 4-6: O2 bioflux . . .is the result of NCP in the mixed layer and will be significantly diminished. . .” do you mean the NCP rate will be diminished, or the estimation of NCP from O2/Ar will be diminished in the presence of vertical mixing. I expect you mean the latter, but it is unclear the way it is currently written.
We mean the latter and clarified this in the text.

9638, line 12: The model has been augmented to predict surface concentrations of gases, which must be dependent on bio/ecosystem model . How is NCP specified in the model? Nutrient supply?
NCP is specified as net production - net consumption of phytoplankton biomass. This was made clearer in the text.

9645, line 17 “We find that BGCCSM generally predicts the meridional variability of ranges in O2 bioflux, suggesting that processes constraining NCP are simulated well” – but if the timing is off are the processes still well simulated?
As earlier mentioned, this is a question about what part of the model we are evaluating. The biogeochemical processes can be well simulated and the models still generate poor results due to problems with the physical model. Errors in timing and regional misfits are examples of problems that could be traced to the physical models due to an the representation of lateral currents, or how the mixed layer dynamics is parametrized. We find it encouraging that the ranges of NCP are represented so well in the models, even when resolved meridionally. This can give us some clues in to when the models can provide robust information.

9645, line 21: Equatorward of 60Sthe models. . . capture the fact that bioflux is seldom <0 – seems like the opposite is true
That’s correct, we changed the text accordingly.

9647, line27, do you mean to say whereas heterotrophic processes?
That’s correct, we changed the text accordingly.

9650, line 5 and following. It seems odd to be presenting this experiment for the first time in the conclusion section; it would be better suited to the ‘discussion’ in section 4.
We agree and changed the manuscript accordingly

Figure 1: report units for NCP and bioflux
We changed the manuscript accordingly

Manuscript prepared for Biogeosciences
with version 5.0 of the LaTeX class copernicus.cls.
Date: 15 December 2014

Evaluating Southern Ocean Biological Production in Two Ocean Biogeochemical Models on Daily to Seasonal Time-Scales using Satellite Chlorophyll and O_2/Ar Observations

Bror F. Jonsson[1], Scott Doney[2], John Dunne[3], and Michael L. Bender[1]

[1]Department of Geosciences, Princeton University, Princeton, New Jersey, USA
[3]Geophysical Fluid Dynamics Laboratory, NOAA, Princeton, New Jersey, USA
[2]Marine Chemistry and Geochemistry Department, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts, USA

Correspondence to: Bror F. Jonsson
(bjonsson@princeton.edu)

Abstract. We assess the ability of ocean biogeochemical models to represent seasonal structures in biomass and net community production (NCP) in the Southern Ocean. Two models are compared to observations on daily to seasonal time scales in four different sections of the region. We use daily satellite fields of Chlorophyll (Chl) as a proxy for biomass, and in-situ observations of ~~and Ar supersaturation~~ O_2 and Ar supersaturation ($\Delta O_2/\mathrm{Ar}$) to estimate NCP. $\Delta O_2/\mathrm{Ar}$ is converted to the flux of biologically generated O_2 from sea to air (~~"O_2 bioflux"~~O_2 bioflux). All data are aggregated to a climatological year with a daily resolution. To account for potential regional differences within the Southern Ocean, we conduct separate analyses of sections south of South Africa, around the Drake Passage, south of Australia, and south of New Zealand.

We find that the models simulate the upper range of Chl concentrations well, underestimate spring levels significantly, and show differences in skill between early and late parts of the growing season. While there is a great deal of scatter in the bioflux observations in general, the four sectors each have distinct patterns that the models pick up. Neither model ~~exhibit a~~exhibits a significant distinction between the Australian and New Zealand sectors, and between the Drake Passage and African sectors. South of 60~~°~~° S, the models fail to predict the observed extent of biological O_2 undersaturation. We suggest that this shortcoming may be due either to problems with the ecosystem dynamics or problems with the vertical transport of oxygen.

~~Overall, the bioflux observations are in general agreement with the seasonal structures in satellite chlorophyll, suggesting that this seasonality represent changes in carbon biomass and not Chl:C ratios. This agreement is shared in the models and allows us to interpret the seasonal structure of satellite chlorophyll as qualitatively reflecting the integral of biological production over time for the~~

~~purposes of model assessment.~~

1 Introduction

Recent years have seen an intense effort to better understand the global biogeochemical cycle. Scientific cruises organized by programs such as CLIVAR, ~~GLODAP~~WOCE, and GEOTRACES have generated a wealth of ~~data, some of which have been aggregated to climatologies of~~ information about physical and chemical tracers in the global oceans~~(e.g. ?) . Field dataand~~ , most of which have been aggregated to climatologies (e.g. ?) . These field data, together with satellite observations of phytoplankton biomass~~have also helped us to assess~~ , have helped in assessing the mean state and variability of the global marine ~~ecosystem~~ecosystems. Concurrently, a number of ocean global general circulation models (OGCMs) with added functionality to simulate biogeochemical processes have been developed, mainly to study trends and ~~variability~~ variabilities in earth's climate and the global carbon cycle.

~~One specific challenge for OGCMs is to simulate vertical exchange of water and tracers in the ocean. Most physical processes that generate vertical mixing and advection act on length-scales several orders of magnitude smaller than the model resolution. Numerical methods with different types of parameterizations have been developed to mimic the effect of such small-scale dynamics (e.g., boundary layer dynamics, diapycnal mixing, and isopycnal mixing). Global climate models are normally~~ This type of global climate model is predominately constructed and tuned to simulate decadal large-scale processes such as the global wind-driven and thermohaline circulation ~~, hydrography and water mass distributions (e.g. temperature, salinity fields), and are optimized to have a high skill in simulating the exchange of water and tracers between the surface ocean and deeper waters below the thermocline (??) . Their skill is evaluated by comparing model simulations with observed global distributions of transient tracers (e.g., radiocarbon and chlorofluorocarbons), especially in the deep sea. The boundary layer parameterizations embedded in OGCMs are also often tested in 1-D against high-frequency observations (e.g. ?) .~~

~~This approach is appropriate for evaluations of model skill on annual to decadal timescales but does not assess how well models simulate underlying processes that act on daily to seasonal scales. An~~ or the distribution of major water masses. However, an increased interest in bio-physical interactions on smaller temporal and spatial scales, together with the recent ability to run global eddy-resolving OGCMs~~have also contributed to the need for determining the skills of 1° resolution OGCMs on these timescales. It is, for example, possible that a model with~~ , have raised the question of how well coarser global climate models perform in daily to seasonal time domains. A model which has a high skill in estimating the basin-wide annual mean phytoplankton biomass ~~does~~ might not correctly simulate process-level details in the seasonal cycle of biological production~~.~~ , such as the onset of spring blooms.

~~Recent studies suggest that high-resolution eddy resolving models have significantly stronger vertical mixing than coarser models. **?** found, for example, that subduction rates were up to 4 times higher in the 1/6° ECCO model of the Southern Ocean than in a model with 1° resolution. Their results also show large variations in vertical mixing and subduction between different areas of the model domain. Another example is a studyby **?** where the investigators modeled the physics and biogeochemistry for a 750 km x 500 km box in the eastern North Atlantic with a 1km resolution model. Their results suggest that subduction is more important than air-sea exchange for removing oxygen from the mixed layer over annual time-scales in this region. Such findings would indicate that the vertical transport of is larger than normally seen in GCMs with lower resolution.~~

~~While these studies both raise important questions about how the models' spatial resolution might affect vertical mixing and advection, they are still model-to-model comparisons. In the analysis, it is also important to include observations in the analysis that reflect the exchange between the mixed layer and the mesopelagic zone on daily to monthly timescales. The observed property must be dynamic in the sense that it has natural sources and sinks large enough to significantly change the concentration on the order of days to week. It is also beneficial if sources and sinks are vertically separated from each other.~~

~~One promising candidate with such characteristics is the ratio of supersaturation between oxygen and argon () . This quantity is influenced by both biological processes on short timescales and vertical transport across the base of the mixed layer, and can hence be used to give a combined evaluation of how well models simulate upper ocean biogeochemical rate processes, sea-air fluxes, and vertical mixing. Such an integrated assessment is valuable since the interaction between physics and biology is a key source for variability on short timescales (days to months). A possible further conversion is to combine with wind data to calculate a flux of biologically generated from sea to air (~~With this study, we aim to do such an evaluation by comparing two state-of-the-art OGCMs with observations of Chl and biological production. Since our main focus is the seasonal cycle and regional variability, we use two properties with high temporal and spatial resolution – satellite derived Chl and in-situ estimations of Net Community Production (NCP) based on the $\Delta O_2/Ar$ method. Chl and NCP are particularly well suited because they represent the ”~~bioflux~~end product” ~~). bioflux is advantageous because the property is the result of net community production (NCP) in the mixed layer and will be significantly diminished by vertical transport of -undersaturated waters into the summertime mixed layer. The following equation expresses the balance of the mixed layer (**?**) :~~ of bottom-up driven biogeochemical models and are both of general interest in climate change studies. Satellite derived Net Primary Production (NPP) is not included this study since it is closely correlated to Chl on the time-scales of interest. We focus on the Southern Ocean south of 40° S since a large number of $\Delta O_2/Ar$ measurements have been collected from this area (**???**) . While satellite derived Chl is a well known and widely used property, $\Delta O_2/Ar$ based NCP will require a more detailed introduction:

$$dO_2/dt = GPP - R + (-FO_2 + D_{in})/h_{ml}$$

~~where dO_2/dt is the time rate of change of dissolved oxygen (units of mol m^{-3} day^{-1}) , GPP is volumetric gross primary production (same units), R is volumetric community respiration rate (autotrophic and heterotrophic), FO_2 is the sea to air gas exchange flux, D_{in} is net input (or loss) of to the mixed layer from ocean physics (i. e., mixing and advection), and h_{ml} is the mixed layer depth. Net community production equals production minus respiration: $NCP = GPP - R$. The oxygen budget of the ocean mixed layer is strongly influenced by exchange with the atmosphere across the air-sea interface.~~

The $\Delta O_2/Ar$ method was developed to estimate oceanic ~~net community production (NCP)~~ NCP by measuring the saturation of ~~and Ar~~ O_2 and Ar in the mixed layer (e.g. **?**). O_2 supersaturation occurs from both biological ~~production and physical processes such as~~ O_2 production and three physical processes: warming, changes in air pressure, and bubble entrainment. It is possible to decouple the biological component from the physical component by using the saturation of ~~Ar, since Ar~~ Ar, since Ar has similar physical properties as O_2 but is biologically inert. Following **?**, we define O_2 supersaturation relative to ~~Ar~~ Ar as:

$$\Delta(O_2/Ar) = \frac{(O_2/Ar)_{sample}}{(O_2/Ar)_{eq}} - 1$$

$$\Delta O_2/\mathrm{Ar} = \frac{(O_2/\mathrm{Ar})_{\mathrm{sample}}}{(O_2/\mathrm{Ar})_{\mathrm{eq}}} - 1 \tag{1}$$

This term is equivalent to the biological O_2 supersaturation. Knowing $\Delta O_2/\mathrm{Ar}$, one can approximate the loss of biological O_2 via gas transfer across the ~~air-sea~~ air–sea interface (hereafter denoted O_2 bioflux) via the relationship

$$O_2bioflux \approx \kappa \cdot \Delta O_2/Ar \cdot O_2eq$$

$$\mathrm{O_2bioflux} \approx \kappa \cdot \Delta O_2/\mathrm{Ar} \cdot \mathrm{O_2eq} \tag{2}$$

where κ is gas transfer velocity based on **?** and ~~O_2eq~~ $\mathrm{O_2eq}$ the concentration of O_2 at equilibrium. It is possible to use either $\Delta O_2/\mathrm{Ar}$ or O_2 bioflux when comparing observations and models. The relative advantages of the two properties are further discussed in Appendix 1; we will use O_2 bioflux in this study.

~~In deriving eq. (3), we have excluded other processes that influence biological supersaturation in~~ One key challenge for OGCMs is to simulate the vertical exchange of water and tracers. Most physical processes that generate vertical mixing or advection act on length-scales several orders of magnitude smaller than the model resolution. Instead, the ~~mixed layer, especially vertical mixing. In high latitudes and during winter, this process generally causes biological saturation to decrease, because is most often undersaturated in waters below the base of the mixed layer during these conditions, whereas Ar is close to saturation. The opposite can be true during the summer for shallow mixed layers, particularly in the subtropics, where a substantial component of NCP (and thus net community production)occurs below the mixed layer, and there is an maximum in the shallow seasonal thermocline (?) . The /Ar supersaturation ishence influenced by the vertical flux of between~~ models use different types of parameterizations to mimic the transport between surface waters and the deep ocean(**??**) . The results are evaluated by comparing model simulations with observed global distributions of transient tracers (e.g., radiocarbon or chlorofluorocarbons), especially in the deep sea. The boundary layer parameterizations embedded in OGCMs are also often tested in 1-D against high-frequency observations (e.g. **?**) . It is, however, less clear how well vertical processes in the mixed layer and ~~deeper waters. is insensitive to the noise in saturation on short timescales from synoptic or seasonal physical processes, such as changes in atmospheric pressure, warming/cooling, and bubble entrainment~~thermocline are represented by these parameterizations. One reason is that, until now, there have been few good methods for evaluating vertical transports on these spatial and temporal scales.

~~The mixed layer supersaturation thus reflects the mass balance between net community production, losses by bioflux, downward transport of from~~ In this study we explore the feasibility of using $\mathrm{O_2}$ bioflux to evaluate how well vertical processes are resolved on shorter timescales. This is possible due to the fact that the actual $\mathrm{O_2}$ balance of the mixed layer is

$$\mathrm{dO_2}/\mathrm{d}t = \mathrm{NCP} + (-\mathrm{FO_2} + D_{\mathrm{in}})/h_{\mathrm{ml}} \qquad (3)$$

where $\mathrm{dO_2}/\mathrm{d}t$ is the time rate of change of dissolved oxygen in units of $\mathrm{mmol\,m^{-3}\,day^{-1}}$, $\mathrm{FO_2}$ is the sea to air gas exchange flux, D_{in} is net input (or loss) of $\mathrm{O_2}$ to the mixed layer ~~to deeper waters, and upward transport of waters with low oxygen concentrations.One consequence of this balance is~~ from ocean physics (i.e., mixing and advection), and h_{ml} is the mixed layer depth.(**?**) Assuming steady state, this equation can be rewritten as a negative relationship between ~~bioflux~~ $\mathrm{O_2}$ bioflux and the net downward vertical transport of ~~, at steady-state~~$\mathrm{O_2}$:

~~$NCP = O_2 bioflux + O_2 vertical flux.$~~

$$\mathrm{NCP} = \mathrm{O_2}\,\mathrm{bioflux} + \mathrm{O_2}\,\mathrm{vertical\ flux}. \qquad (4)$$

~~Most studies of mixed layer aim to constrain NCP from bioflux estimates. In some studies, the vertical fluxes were estimated and equation (4) was used to calculate NCP. In other studies, the vertical flux of biological was assumed to be negligible when calculating NCP (e.g. ??) . The resulting bias in NCP has been diagnosed using models where both NCP and bioflux are prescribed (?) . ? found that the influence of biology and physics on bioflux differed between two ocean models, reflecting choices in physical and biological parameterizations as well as surface forcing and, to some degree, model resolution. Here we compare the spatial and temporal patterns of bioflux in the same two models with the distribution observed in the Southern Ocean, with the hope that the use of multiple models will highlight both commonalities and model-specific skill and errors.~~

~~Our focus in this study is the Southern Ocean south of 40°S. This is a region of critical importance for the global biogeochemical cycle and for the uptake of anthropogenic carbon- two processes connected to the vertical exchange of water between the shallow and deep parts of the ocean. The Southern Ocean is also a key region for the global thermohaline circulation since here deep water is upwelled to the surface, transformed by biological activity and air-sea exchange, and eventually exported as intermediate, mode, or bottom waters. Another benefit with focusing on this region is that a large number of measurements have been collected from this area. In this study, we use measurements from several cruises between 1999~~ O_2 bioflux can hence be used to give a combined evaluation of how well models simulate upper ocean biogeochemical rate processes, sea–air O_2 fluxes, and ~~2011 that have been reported in ??? . We compare this dataset with two state-of-the-art global ocean general circulation models which both have the added functionality to simulate bioflux (?) . The two models have coarse, non-eddy resolving spatial resolutions (1-3°) and are forced by CORE2 synoptically-varying~~vertical mixing. Such a combined assessment is valuable since the interaction between physics and biology is a key source for variability on short timescales (days to months).

We will first compare regional and seasonal patterns in Chl and O_2 bioflux between observations and models, continue with exploring how O_2 bioflux can show discrepancies in how the models simulate vertical transports, ~~re-analyzed winds and atmospheric properties.~~

~~In this paper, we explore the difference between and bioflux, compare our observations with model output,~~ and finally discuss ~~what the results tell us about net community production and the summertime exchange between the mixed layer and the mesopelagic~~how the results could help to identify mechanisms that contribute to mismatches between observations and models.

2 Methods

~~This evaluation is based on combining data from observations , models, and remote sensing~~

The main focus of the study is to compare model output with in-situ observations and satellite derived properties. The data from each source have different temporal and spatial resolution and span over different ranges in time. To compensate for these discrepancies, we re-grid in-situ observations and satellite fields to a model grid with roughly 1° resolution at equal intervals in time. We also combine observations from different years but the same year day (e.g. ~~2003–01–01, 2004–01–01, 2005–01–01~~1 January 2003, 1 January 2004, 1 January 2005) to a climatological year with a daily time resolution.

2.1 Satellite data

We use remotely observed chlorophyll concentrations from MODIS/Aqua on the Level-3 ~~9 km x 9 km~~ 9 km × 9 km grid. Daily satellite images from 2003 to 2010 were aggregated to the model grids and averaged to a year-day climatology. Satellite data, particularly in the Southern Ocean, suffers from a high frequency of days where clouds, light conditions, sea-ice, or other problems disqualify the observations. The relative frequency of ~~such days for~~ days with valid information in our dataset is between 5~~%~~ % and 15~~%~~ % on the original grid and between 20~~%~~ % and 50~~%~~ % on the aggregated model grid.

2.2 ~~In-Situ ObservationsWe use~~ In-situ observations

We use $\Delta O_2/Ar$ observations from 19 Southern Ocean cruises between 1999 and 2009, all occurring during the austral summer. The geographical locations of the respective ship tracks are presented in ~~figure~~ Fig. **??**. The measurements were conducted by two different methods: water was collected in bottles and analyzed in lab (shown as blue in ~~figure~~ Fig. **??**) on 16 cruises, and $\Delta O_2/Ar$ was measured directly using a ship-borne flow-through system in 3 cruises (shown as red in ~~figure~~ Fig. **??**). The measurements are clustered in space and time reflecting tracks of the ships of opportunity used in this study. We use these sampling clusters in our analysis as natural areas to compare and contrast different parts of the Southern Ocean. A more detailed description of the sampling strategies~~and measurement methods~~, measurement methods, and data sources can be found in ~~???~~ **?** and **??**.

Bioflux is calculated as the product of the biological O_2 supersaturation and the gas transfer velocity. The latter is determined using the **?** parameterization expressing gas transfer velocity in terms of a quadratic function of wind speed and the Schmidt number. We do the calculation using daily averages of the NESDIS wind product with a 0.5° resolution based on data from the QuikSCAT satellite (**?**). The gas transfer velocity for each $\Delta O_2/Ar$ measurement is calculated from the daily-mean local wind speeds during the 60 days preceding collection. A time-weighted value for the gas transfer velocity is calculated based on the fraction of the mixed layer flushed in each subsequent interval until sampling (**??**). The resulting gas transfer velocity is then used in ~~equation~~ Eq. (3~~to calculate bioflux from~~) to calculate O_2 bioflux from $\Delta O_2/Ar$ supersaturation. A detailed analysis of possible uncertainties affiliated with the method can be found in **?**.

2.3 ~~Models~~

Finally we use Mixed Layer Depth (MLD) as a baseline diagnostic of how well vertical physical processes in the surface ocean are resolved. A total of about 75000 vertical profiles from Argo floats between 2001 and 2012 are used to estimate in-situ MLDs for the different regions. The observed MLD is defined as the depth where density is 0.03 $\mathrm{kg/m^3}$ higher than at the most shallow observation.

2.3 Models

The observations are compared with output from the ocean components of two IPCC-class ocean biogeochemical models. The TOPAZ ocean model is built upon version 4 of the modular ocean model (MOM4 ?) with a vertical z-coordinate and a horizontal B-grid with a tri-polar coordinate system (North America, Siberia, and Antarctica) to resolve the Arctic. The model has a nominally 1-degree horizontal resolution globally with higher meridional resolution near the equator (to 1/3~~degrees~~ °). There are 50 vertical layers; resolution is 10~~m~~ m in the upper 200~~m~~ m, and coarser below. MOM4 includes a representation of the k-profile parameterization (KPP) planetary boundary layer scheme (?), ~~Bryan-Lewis~~ Bryan–Lewis deeper vertical mixing, ~~Gent-McWilliams~~ Gent–McWilliams isopycnal thickness diffusion (?), bottom topography represented with partial cells, isotropic and anisotropic friction, and a~~multiple dimensional flux limiting~~ multiple-dimensional-flux-limiting tracer advection scheme using the third order Sweby flux limiter. For these studies, the ocean model is forced by prescribed boundary conditions from the reanalysis effort of the ECMWF and NCAR Common Ocean-ice Reference Experiments (CORE).

BGCCSM is based on the Los Alamos National Laboratory Parallel Ocean Program (POP) (?). In our application, the grid is symmetric in the ~~southern hemisphere with a~~Southern Hemisphere with a zonal resolution of 3.6~~°~~°. Meridional resolution decreases from 1.8~~°~~° at mid-latitudes to about 0.8~~°~~° at high and low latitudes. The surface layer is 12~~m~~ m thick; there are in total 5 layers to 111~~m~~ m, and 25 layers to the bottom. This model invokes ~~Gent-McWilliams~~Gent–McWilliams' isopycnal mixing and the KPP upper ocean model, and is also forced by prescribed boundary conditions from CORE.

The ~~air-sea fluxes of and~~ air–sea fluxes of O_2 and CO_2 in both models are computed using prescribed atmospheric conditions (surface pressure, mole fraction), model-predicted surface water concentrations, NCEP surface winds, and the quadratic dependence of the gas exchange coefficient on wind-speed ~~model of~~ (?). Argon was added as a prognostic tracer to the simulations in both models in an analogous fashion to O_2; i.e., O_2 and Ar solubility are similarly determined using model temperature and salinity, and ~~Ar~~ Ar uses the same gas-exchange parameterization as O_2.

Both models include complex biogeochemistry/ecosystem components with macro- and micronutrients, organic matter, and three phytoplankton functional groups: small phytoplankton, large phytoplankton, and diazotrophs. Both invoke co-limitation by iron, light, and nitrogen, mediated in part by their influence on the Chl~~:~~ : C ratio, and slower rates of photosynthesis at lower temperatures. Diazotrophs have high N~~:~~:P ratios and low photosynthetic efficiencies (TOPAZ) or high iron requirements (TOPAZ, BGCCSM). NCP is calculated as the difference between production and consumption of carbon by the different functional groups.

It is possible to use either ~~or~~ $\Delta\mathrm{O_2/Ar}$ or $\mathrm{O_2}$ bioflux for comparing observations with models. One could argue that ~~is a~~$\Delta\mathrm{O_2/Ar}$ is a more robust property since it is an observed quantity, whereas $\mathrm{O_2}$ bioflux is a derived product that depends on the wind field. On the other hand, model $\Delta\mathrm{O_2/Ar}$ depends on the choice of wind forcing, whereas $\mathrm{O_2}$ bioflux is assumed to be ~~mainly controlled by~~ directly linked to NCP. Our tests show that both methods give similar accuracy and we choose to use $\mathrm{O_2}$ bioflux since its units are more appropriate for the current study (~~See appendix~~ see the Appendix for further discussion).

3 Results

First we test the feasibility of aggregating data from different years into a single climatological year. This approach is useful only if the difference between seasons is significantly larger than the interannual variability. We test the assumption by comparing how model $\mathrm{O_2}$ bioflux changes from one date to another in a climatological year, against the standard deviation at the same dates between a number of years. ~~Fields~~ As a test case, fields from BGCCSM were used to generate climatologies for the days ~~November 15th and December 15th~~ 15 November and 15 December by averaging data for the entire Southern Ocean from four consecutive model-years. Our results show that the mean difference is 19.8~~mmol~~ $\mathrm{mmol\,O_2\,m^{2}\,d^{-1}}$ between the two different dates, whereas the standard deviation over four years on the respective dates is only 8.7~~mmol~~ $\mathrm{mmol\,O_2\,m^{-2}\,d^{-1}}$. It is hardly an unexpected result that spring values of ~~bioflux differs~~ $\mathrm{O_2}$ bioflux differ from summer since the Southern Ocean is a high-latitude region. This result encourages us to aggregate observations and model-data from different years into a climatological year.

Next, we compare the model simulations with observations. Chlorophyll simulated by BGCCSM and TOPAZ is related to satellite observations retrieved by the MODIS/Aqua mission from ~~2003-01-01 to 2010-12-31.~~ 1 January 2003 to 31 December 2010. Each daily satellite image is reprojected from a~~native 9x9km~~ native 9 km × 9 km resolution to the TOPAZ' ~~1°x1°~~ 1° × 1° grid and the fields are aggregated to a climatological year, as mentioned earlier. The resulting datasets are analyzed in four geographical sectors, shown as shaded areas in ~~figure~~ Fig. **??**. We zonally average the data in each sector to a~~Hofmø~~ Hofmøller diagram with latitude on the y-axis, time on the x-axis, and daily zonal Chl averages as colors (~~figures~~ Figs. **??**–**??**). It should be noted that one problem with using satellite-retrieved Chlorophyll is a systematic lack of data during winter due to low light conditions and sea-ice cover in combination with the satellite's track. These periods are shown as gray areas in the figures. Figure **??** show a more detailed representation of the data at 60° south.

Figures **??**–**??** present model NCP and bioflux vs. our observations. Panel ~~A~~ a in each figure shows the temporal evolution of NCP ~~versus~~ vs. latitude in the two models, and panel ~~B~~ b the corresponding model O_2 bioflux. The sampling locations are indicated on the model plots by gray circles. Panel ~~C~~c, finally, presents O_2 bioflux from the observations shown in ~~figure~~ Fig. **??**. As mentioned earlier, all measurements that fall on the same year-day and grid cell are combined to one mean value, which is indicated by the color of the dot. Note that the aggregated values of observed O_2 bioflux presented together with TOPAZ are somewhat different from the ones connected to BGCCSM, since the two model grids are different.

A general pattern arises where the models simulate upper ranges of Chl well but underestimate spring levels significantly. ~~Both models also show~~ Each model also shows differences in skill between early and late parts of the growing season, performing better during the spring and early summer. While there is a great deal of scatter in the bioflux observations, the four sectors each show distinct patterns, ~~wheras~~ whereas the models exhibit little distinction between the Australian and New Zealand sectors, and between the Drake Passage and African sectors. Next, we compare the simulated fields and data for each of the study regions in more detail.

3.1 Drake ~~Passage~~passage

The seasonal change in satellite-derived Chlorophyll for the Drake Passage sector is shown in ~~Figure **??**, panel A~~ the top panel of Fig. **??**. It is possible to discern a seasonal cycle, even ~~if~~ though we lack winter and early spring data in higher latitudes. Towards ~~to~~ the South, the earliest retrieved concentrations are between 3 and 5 times or more lower than maximum values at late spring/summer peak. Both the magnitude and timing of maximum summer Chl concentrations vary with latitude, with the strongest and latest blooms occurring in high latitudes. [Chl] reaches 1~~mg m-3~~ $\mathrm{mg\,m^{-3}}$ south of 70~~°~~° S, decreases to 0.5~~mg Chl m-3~~ $\mathrm{mg\,Chl\,m^{-3}}$ between 60~~°~~° S and 70~~°~~° S, and increases again north of 50~~°~~° S. The growing season, as inferred from the period of elevated summer Chl, is about three months south of 70~~°~~° S and generally lengthens going to the north. The ~~band~~ bands of very high Chl north of 45~~° S is~~° S and south of 70° S are likely due to transport of sedimentary iron from coastal waters generating elevated biological production.

Panels ~~B and C~~ b and c present Chl climatologies from TOPAZ and BGCCSM. The grey shadings indicate where satellite coverage is missing. It is clear that the seasonal chlorophyll progression in both models diverge from observations, whereas the magnitude of peak Chl concentrations is simulated rather well. TOPAZ underestimates Chl concentrations somewhat early in the season and generates an intense spring bloom extending over the entire sector from 75~~°~~° S to 40~~°~~° S. (~~”Bloom”~~ “Bloom” as used here indicates a transient period of high [Chl] lasting no more than 1 month.) In the observations high [Chl], once established, continues throughout most of the summer, and concentrations ~~> 1 mg m-3~~ $> 1\,\mathrm{mg\,m^{-3}}$ are limited to the region south of 70~~°~~° S. In these southern regions, the onset of the TOPAZ bloom comes ~~1-2~~ 1–2 months earlier than in the observations. The spring bloom in TOPAZ collapses after about a month with too low concentrations of Chl over the austral summer as a result. Finally, the season ends with TOPAZ generating a weaker fall bloom during part of the season without satellite coverage. South of 50~~º~~° S, BGCCSM begins the season with much lower Chl concentrations than both TOPAZ and satellite observations. The model generates a much more intense bloom between 60~~º~~° S and 70~~º~~° S than observations suggest, but has good skill in predicting the timing and magnitudes south of 70~~º~~° S. Both models suggest that during the bloom, biomass is significantly higher in the frontal regions compared to observations.

O_2 bioflux observations in the Drake Passage region originates both from transects crossing the Drake Passage and nearby cruises in the Palmer Long Term Ecological Research program annual survey west of the Antarctic Peninsula. Our results are thus partly influenced by coastal processes outside the models' domain. In the Drake Passage sector, TOPAZ and BGCCSM show similar patterns in NCP (panel ~~A~~a) north of 65~~º~~° S. South of 65~~º~~° S, TOPAZ simulates intensive NCP over the summer, whereas BGCCSM has close to zero net production. Both models have predominantly negative O_2 biofluxes south of 50~~º~~° S in October (panel ~~B~~b), and a progressive change to positive fluxes from north to south with time. The negative values of O_2 bioflux in TOPAZ are about twice as large as they are in BGCCSM. In TOPAZ, O_2 bioflux is positive throughout the domain in January, whereas in BGCCSM there is a southerly region of negative flux throughout the season.

Panel ~~C of figure~~ c of Fig. **??** shows observations of O_2 bioflux in the region. The pattern of negative fluxes south of 60~~º~~° S in November and December corresponds qualitatively with the models. There is a wealth of observations south of 60~~º~~° S in January and February, but they show significant variability. Some observations suggest biological production rates of up to 50~~mmol~~ $\mathrm{mmol\,m^{-2}\,d^{-1}}$, which is similar to the levels predicted by TOPAZ. Others correspond to the results from BGCCSM with O_2 bioflux estimates near or below zero. The mean observed O_2 bioflux values in January and February south of 65~~ºS are 8 ± 20 mmol~~° S are $8 \pm 20\,\mathrm{mmol\,O_2\,m^{-2}\,d^{-1}}$. The corresponding values for BGCSSM are ~~9 ± 13~~9 ± 13, and for TOPAZ, ~~16 ± 27.~~ 16 ± 27. The observations in April suggest undersaturated conditions that correspond with BGCCSM but not with TOPAZ.

3.2 South Africa

The observations in the South Africa sector (~~figure **??**, panel A~~Fig. **??**, top panel) show similar patterns to the Drake Passage sector for satellite-derived Chl, with the main exception of a weaker bloom north of 50~~º~~° S. Note the lack of observations south of 70~~º~~° S where of Chl concentrations may be very high. TOPAZ (panel ~~B~~b) begins the season with somewhat higher concentrations of Chl than observed and, with time, generates a significantly stronger bloom. As in the Drake Passage, the TOPAZ bloom crashes mid-summer, after which model Chl concentrations are significantly lower than observations. BGCCSM starts the season with much lower Chl concentrations than the satellite data and generates an exaggerated bloom as well, especially south of 55~~º~~° S.

NCP and O_2 bioflux climatologies (~~figure~~ Fig. **??**) show that TOPAZ has an earlier and more intense band of positive NCP than BGCCSM (panel ~~A~~a). BGCSSM has a slow southward progression of the the positive NCP band with NCP reaching 20~~mmol~~ mmol m^{-2} d^{-1} at 40~~º~~° S in October and at 65~~º~~° S in February. TOPAZ, on the other hand, has a more uniform pattern in which most of the region reaches these levels of NCP by mid-November. The two models have O_2 bioflux patterns (shown in panel ~~B~~b) that vary in a fashion similar to that of the Drake Passage region. Both models start with negative ~~bioflux: <-30~~ O_2 bioflux: < -30 m^2 d^{-1} in TOPAZ and ~~∼ -10 mmol~~ ~ -10 mmol m^{-2} d^{-1} in much of the domain in BGCCSM. This negative bioflux switches to positive values faster in TOPAZ than in BGCCSM, most likely due to the earlier onset of biological production. Finally, BGCCSM has a slightly earlier and stronger switch to negative O_2 bioflux at the end of the growing season than does TOPAZ.

We have O_2 bioflux observations from four crossings in the South Africa sector (~~figure ??panel C~~Fig. ??) distributed in time from December to late March. The November-December transect has a pattern of weakly positive O_2 bioflux in the north that turns negative below 50~~º~~° S. Both models show a somewhat different pattern with positive O_2 bioflux south of 50~~º~~° S as well. While both models show positive values, the observed January transect has predominantly negative biofluxes. Finally, the two transects in March show a pattern with high positive O_2 bioflux in the north and significant negative bioflux at 55~~º~~° S. BGCCSM has a better skill in recreating these patterns than does TOPAZ.

3.3 Australia and New Zealand ~~Regions~~regions

~~The~~ Chl in the Australian and New Zealand sectors (~~figure~~ Figs. ?? and ??) follow the general pattern of Drake Passage and South Africa. The main exception is BGCCSM generating spring blooms further north and in a spottier pattern than in the sectors discussed earlier. In this model, the New Zealand sector has four distinct areas/periods of high Chl concentrations: in ~~Jan-Feb~~ January-February south of 70~~°S, Nov-Dec~~ ° S, November-December between 60~~º~~° S and 70~~°S, Jan~~ ° S, January between 45~~º~~° S and 55~~º~~° S, and ~~Oct-Nov~~ October-November north of 45~~º~~° S. This peculiar pattern could be due to interactions with hydrology in the frontal regions. The Australian sector show similar patterns for the latter three areas whereas the region south of 70~~º~~° S lack data. TOPAZ shows indications of being out of phase with both MODIS and BGCCSM after the beginning of December, whereas BGCCSM captures the seasonal cycle in MODIS better.

NCP climatologies from the Australian sector (~~figure ??, panel A~~Fig. ??, top panel) show that TOPAZ generates short intense periods of positive NCP early in the growing season. In the southern reach of the domain, NCP stays high into the fall whereas elsewhere summer and fall NCP values are low (-10~~mmol~~ mmol m^{-2} d^{-1}). BGCCSM, on the other hand, has a much longer and more intense period of positive NCP than in other parts of the Southern Ocean. In this model, spring conditions south of ~~45-50°~~45–50° S are dominated by weaker negative O_2 biofluxes when compared to other regions in both models (panel ~~B~~b). BGCCSM switches to negative bioflux at the end of the summer whereas O_2 bioflux in TOPAZ stays positive throughout March.

The observations in this part of the ocean differ from those simulated by TOPAZ. For example, almost all measurements north of 55~~°~~° S report higher ~~biofluxes, > 25~~ O_2 biofluxes, > 25 $\mathrm{mmol\,m^2\,d^{-1}}$, in February. As well these field data exhibit low or even negative biofluxes at 65~~°~~° S where TOPAZ predicts strong biological production and high O_2 bioflux from mid-November to mid-February. BGCCSM is more in line with observed patterns, with the main exception of an early onset of negative O_2 bioflux in the fall at 40~~°S -50°S~~ ° S–50° S in the model where the observations still suggest strong positive biofluxes. In general, between 50 and 60~~°~~° S, TOPAZ and BGCCSM both ~~simulates a~~simulate a strong spring bloom whereas observations show a simpler patters of sustained high production throughout the latter part of spring and summer. Model NCP and O_2 bioflux in the New Zealand region are similar to values in the Australian sector (~~figure ??, panels A and B~~Fig. ??). The O_2 bioflux observations in panel C show considerable scatter but in general are marked by high values north of the Polar front at 60~~°~~° S, with low or negative values of bioflux to the south at most times. Both models capture the highly negative values of bioflux early in the growing season south of ~~60-65°~~60–65° S, and occasionally negative values of bioflux later in the growing season. BGCCSM simulates our observations of sustained high NCP north of the Polar Front from November through February.

3.4 Cross ~~Regional~~regional, Southern Ocean ~~Analysis~~analysis

The data-model comparisons of O_2 bioflux for the different regions have certain patterns in common. In both models, the spring period of strong positive flux starts later in high latitudes. In general, TOPAZ tends to have an earlier, shorter, and more intense period of high Chl concentrations, than does BGCCSM. Observed values of O_2 bioflux are much more variable than simulated values. This is expected given the relatively smooth and coarse fields of the models, and likely reflects the absence of mesoscale processes in the models.

Our next step is to compare the seasonal range in O_2 bioflux values between observations and models as a function of latitude (~~Figure~~ Fig. ??). We aggregate the observations to year-days and model grid-cells as described above, but compare the resulting values with corresponding individual data points in the models matched by both location and collection time. Such comparison using individual data points suffers more from small-scale spatial mismatches than the zonal model averages used earlier, but allows us to better compare the seasonal range of O_2 bioflux values between models and observations. Figure ?? shows ~~bioflux versus~~ O_2 bioflux vs. latitude for the two models and observations in each of the previously defined regions. We find that BGCCSM generally predicts the meridional variability of ranges in O_2 bioflux, suggesting that processes constraining NCP are simulated well. The seasonal maximum of model O_2 bioflux might not occur at the same time in the models as in the real world, but the magnitude of the maximum values seems to be fairly well predicted. Equator-ward of about 60~~°~~° S the models also tend to capture ~~the fact that biofluxis seldom < 0, reflecting~~ range and meridional structure of negative O_2 bioflux, which reflects both low wintertime NCP and physical transport. In contrast, the models do not capture well the observed strong negative O$_2$ bioflux at high latitudes. TOPAZ also tends to exaggerate high positive O$_2$ bioflux in some areas, such as 60~~°S – 70°S~~° S–70° S in the Drake Passage and New Zealand regions, whereas O$_2$ bioflux is underestimated in TOPAZ between 40~~°~~° S and 50~~°~~° S in the Australian and New Zealand regions.

~~Finally, we~~ We compare observed and simulated O$_2$ bioflux values for the same locations and times. The scatter plot in ~~Figure ?? compares~~ Fig. **??** compares O$_2$ bioflux simulated by TOPAZ (blue) and BGCSSM (red) with observations binned to the same year-days on the respective model's grid. It is clear from this figure that both models show a low correlation with the observations (~~=0.024~~ $r^2 = 0.024$ for TOPAZ and ~~=0.23~~ $r^2 = 0.23$ for BGCCSM). This low correlation is expected: lags in time or displacements in space can generate large differences between the models and observations even if the fundamental processes are simulated with high skill (**?**). Further, the field data contain mesoscale variability that cannot be captured in the models (even if the models were eddy-resolving, the details of the simulated turbulent fields would differ from observed). It is also clear that the distributions of model-data residuals (difference from a~~diagonal 1:1~~ diagonal 1 : 1 line) are asymmetrical. The upper right quadrant, where both observations (obs) and model data (mod) are positive, has considerable scatter about the ~~1:1~~ 1 : 1 line but no apparent bias. The lower right quadrant (negative model, positive observation) is mainly empty, showing that the model rarely predicts undersaturation when the observations report supersaturation. An exception is the clustering of TOPAZ NCP values about zero, a consequence of low production simulated by that model after an intense bloom. The upper left quadrant is heavily populated, showing that the models frequently simulate positive values of bioflux when the ocean is in fact undersaturated. Consistent with this pattern, the results in the lower left quadrant show that, when data and models agree that bioflux is negative, the observations are more negative than the models. The overall picture is that both models have a positive bias in predicting bioflux, mainly due to fewer negative values compared to the observations (~~figure~~ Fig. **??**). For all data points the model bias is statistically significant for both a~~paired t-test~~ paired t test (BGCCSM: ~~n=271, t=-4.803, p=0.000~~$n = 271$, $t = -4.803$, $p = 0.000$; TOPAZ: ~~n=273, t=-2.870, p=0.004~~$n = 273$, $t = -2.870$, $p = 0.004$) and a one-tailed binomial test (BGCCSM: pos~~=167, tot=271, p=0.00005~~ = 167, tot = 271, $p = 0.00005$; TOPAZ: pos~~=151, tot=273, p=0.035~~ = 151, tot = 273, $p = 0.035$). Neither model shows a statistically significant bias when only positive values are considered.

4 ~~Discussion~~

3.1 Mixed Layer Depths

Finally, we create MLD climatologies for the different regions by integrating Argo and model MLDs using earlier mentioned methods (Figure **??**). Both models are able to simulate the general trends rather well, with deep mixed layers in winter, a soaring in proving and shallow mixed layers during the summer. Regional structures are also similar in general, with the main exception of deep winter mixing extending too far south in TOPAZ, even if this is somewhat inconclusive due to lack of observations. The two main differences between models and observations are that the spring shoaling tend to be later and more gradual in the observations. There is also much more small-scale variability in the Argo climatology, which could be explained by sparse data but also that the models have a too smooth mixed layer dynamics.

4 Discussion

When comparing model and satellite climatologies of Chl concentrations and O_2 bioflux, we find both similarities and significant differences. The models are able to predict spring and summertime maximum levels of Chl and O_2 bioflux well, but ~~generate ~~levels are much too low ~~values of Chl in the onset of the springseason~~during the winter and early spring. Such underestimations are particularly important in the case of Chl, which by nature has a lognormal distribution (**?**) and is hence skewed towards low values. One explanation for this behavior is a combination of model grazing and/or phytoplankton mortality being too strong in the winter, as reported for the BGCCSM in the subpolar North Atlantic by ~~Behrenfeld et al (2013)~~**?** . These patterns could also be explained by the two models simulating too weak vertical export of phytoplankton during summer and too strong export during winter. **?** has shown that MODIS/Aqua generally underestimate the dynamical range of Chl in the Southern Ocean, which would suggest that these differences might be even larger.

Another general pattern is that the models tend to simulate the increases in Chl ~~duing ~~during spring and early summer rather well, but show highly diverging behavior later in the season. In TOPAZ, the ecosystem tends to crash in early January with much too low biomass as an effect, whereas BGCCSM has patches where far too much Chl is produced. ~~This difference between early and late parts of the summer season can ~~Work by (**?**) suggest that such differences in skill over the season could be explained by ~~which processes ~~changes in processes that control the ecosystem~~behavior at different times~~. The spring bloom onset is thought to be controlled mainly by physical factors such as light, temperature, and vertical stratification, ~~and ~~whereas the summer peak magnitude ~~is~~ to be controlled by nutrient availability, grazing, mortality, and other ecological and biogeochemical factors~~(Hashioka et al. , in press). The ~~. Finally, the decrease of Chl concentrations after a~~summer peak depends mainly ~~ summer peak mainly depends on ecosystem dynamics such as grazing, succession, and other interactions between organisms, as well as vertical export of particulate organic matter. Phytoplankton production ~~are ~~is in general better resolved by the models, whereas ~~autotrophic ~~heterotrophic processes and vertical transports are challenging to implement well~~to the models. There ~~. These processes are often stochastic in their behavior and there is a lack of observation to parameterize them accurately.

~~Even if the two models generally predict summer season ranges of bioflux~~ With respect to models and observations, the misfit in spring Chl biomass is likely explained by the differences in mixed layer dynamics between models and observations. Too deep consistent winter mixed layers followed by a rapid shoaling without much small-scale variability , as observed in the models, would lead too low winter biomass and exaggerated spring blooms, again as observed. Differences in the timing of mixed layer shoaling in the spring, how the mixed layer shoals and short time variability, all can strongly affect the onset of the spring bloom since the availability of light and nutrients in the mixed layer will be impacted. Problems with MLD dynamics has been shown by earlier studies in climate models similar to TOPAZ and BGCCSM (e.g. **?**, and references within) .

Both models show significant skill in simulating how the ranges of O_2 bioflux vary meridionally in the Southern Ocean north of 60~~°Swell, there are~~° S. The fact that the models provide such reasonable results suggests that the models constrain ecosystem processes in the region rather well. The models show, however, large regional variations ~~. The differences~~ that tend to follow the earlier discussed patterns in Chl concentrations~~discussed earlier~~, both when ~~comparing the two models~~ compared with each other and ~~when comparing the models~~ with observations. We also find different and varying patterns in the seasonal cycle of biological net community production ~~,~~ with TOPAZ often having a shorter, more intense, growing season than BGCCSM. That ecosystem processes seem to be well simulated, but the models still have regional and timing issues, could suggest problems with the physical model, such as how well lateral currents are represented or how mixed-layer and thermocline dynamics are parameterized.

One important difference between models and observations is that models fail to predict observed events of negative O_2 bioflux in waters south of 60~~°~~° S. We suggest two possible explanations for why the models lack these events: problems with the ecosystem dynamics and problems with the vertical transport of oxygen. It is possible that observed summertime undersaturation is generated by net heterotrophy (negative NCP) temporally decoupled from earlier biological production. O_2 supersaturation from periods of positive NCP would then be lost via ~~air-sea~~ air–sea exchange before the start of ~~net-respiration~~net respiration. It is also necessary for particulate organic carbon to remain long enough in the mixed layer to be respired, and hence particle export has to be limited. Such events have been observed during the GasEx III experiment ~~(Hamme et al., 2012)~~ (**?**) but are not generated by either model. Both TOPAZ and BGCCSM simulate NCP to evolve more smoothly than observations across time and space, even if significant variability is seen ~~(Jonsson et al. , 2013).~~ **?** .

Our second explanation as to why models fail in capturing summertime mixed layer O_2 undersaturation is that they underestimate rates and characteristics of vertical mixing. Several studies have shown the potential for entrainment and submeso-scale processes to transport waters between the thermocline and the mixed layer on short time scales. Such events would introduce O_2 undersaturated waters into the mixed lay and generate the type of conditions we find in our observations.

OGCMs such as TOPAZ and BGCCSM lack the spatial resolution to resolve these kinds of processes and hence the ability to generate the undersaturated conditions we observe.

We cannot conclusively disprove either explanation for the models' failure to produce negative O_2 bioflux in summer but ~~other studies support the suggestion that vertical transport is an important factor. Simulations by ?? , both~~ both the mismatches in MLD dynamics and the specific patterns in O2 undersaturation strongly suggests that vertical transports is the dominating factor. This explanation is also supported by other studies such as ?? . Both find much stronger subduction rates south of 60~~º~~° S when comparing high resolution models to ones with lower resolution~~such as TOPAZ and BGCCSM~~. The areas of increased subduction described by Sallee and Rintoul correspond very well with the latitude bands where we observe undersaturated conditions not generated by the models. (Below 60~~º~~° S for the New Zealand, Australia, and Drake Passage sections; Between 60~~º~~° S and 40~~º~~° S for the South African section.) ~~.~~Other studies furthermore suggest that such physical processes ~~enabled~~ captured by higher spatial resolutions tend to be episodic in nature, making them prime candidates to generate the undersaturated O_2 conditions seen in our observations.

One might argue that the lower frequency of negative O_2 bioflux in the models is simply due to model overestimation of NCP. Such a conclusion is not consistent with the models' accurate simulations of the seasonal range of positive O_2 bioflux (cf. ~~figure 11~~fig. ??). The models appear to selectively overestimate O_2 bioflux when the observations of O_2 bioflux are negative.

5 Conclusions

~~Both the~~ The fact that both TOPAZ and BGCCSM ~~models simulate upper ranges of~~ simulate the ranges of NCP and peak summer Chl concentrations well ~~but underestimate spring levels significantly. They also show differences patterns in their skill during early and late parts of the growing season~~suggest that the models are able to parametrize at least physiological ecosystem processes depending on mechanistic relationships rather well. We find ~~a great deal of scatter overall between simulated bioflux~~ three main problems in the models~~and observed values when the models are sampled in a analogous fashion to the observations. There are, however, distinct patterns in each sections that both models replicate. and the two models generally predict summer season ranges of bioflux in the Southern Ocean north of 60°S well. South of 60°S, the models fail to predict the observed extent of biological undersaturation. We suggest that this shortcoming may be due either to problems with the ecosystem dynamics or problems with the vertical transport of oxygen.~~: Errors in the timing and initial biomass levels of the spring bloom, regional displacements of high biomass and NCP, and failure to generate observed extents of negative O_2 bioflux. The combined picture of how models and observations compare, together with results from other studies, suggest that the main reasons for these errors are how vertical physical processes are simulated.

~~To evaluate or~~

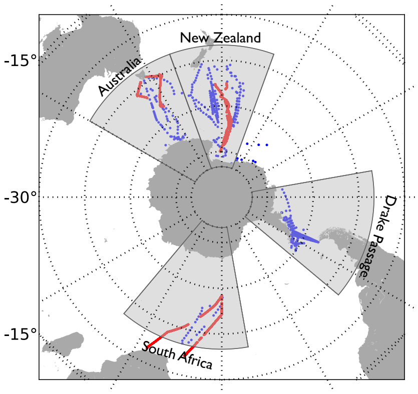
Acknowledgements. This work was supported in part by funding from National Aeronautic and Space Administration (NASA NNX08AF12G) and National Science Foundation (NSF OPP-0823101).

6 Appendix

To evaluate $\Delta \mathrm{O_2/Ar}$ or $\mathrm{O_2}$ bioflux for comparing observations with models, we use a box model that calculates the evolution of ~~and~~ $\Delta \mathrm{O_2/Ar}$ and $\mathrm{O_2}$ bioflux from prescribed time-series of NCP and wind. The box is a 50~~meter~~ m deep water column with no vertical or horizontal advection or mixing. ~~air-sea~~ $\mathrm{O_2}$ air–sea exchange is simulated using $\mathrm{O_2}$ saturation and the **?** gas transfer velocity parameterization; the exact setup is described in **?**. We conduct three experiments with the box model using a time series of NCP from TOPAZ at 160° W, 61°~~S (figure ??, panel A~~ S (fig. **??**a). In the first experiment, represented by the blue lines in panels ~~B-D~~b–d, we use CORE-2 atmospheric reanalysis winds with a~~2x 2~~ $2^\circ \times 2^\circ$ resolution from the same position as the NCP observations (panel ~~B~~b). The resulting ~~supersaturation and~~ $\Delta \mathrm{O_2/Ar}$ supersaturation and $\mathrm{O_2}$ bioflux are presented in ~~Panels C and D~~ panels c and d respectively. In the second experiment, which is presented with red lines in ~~Panels B-D~~panels b–d, we instead use Quickscat satellite-derived wind data with a~~0.25x 0.25~~ $0.25^\circ \times 0.25^\circ$ resolution from the same location as the NCP, to simulate the effect of a different wind product. When comparing the two cases, we find that $\mathrm{O_2}$ bioflux is a more robust indicator of the underlying ecological behavior (NRMSD~~=5.0%) than (NRMSD=10.7%~~ $= 5.0\,\%$) than $\Delta \mathrm{O_2/Ar}$ (NRMSD $= 10.7\,\%$), which we find more sensitive to short-term differences in the wind forcing applied to the model.

This effect, however, is only true if $\mathrm{O_2}$ bioflux is calculated using the same wind history as that used to generate $\Delta \mathrm{O_2/Ar}$ in the box. In the real world, the forcing is the true wind stress. However, ~~Bioflux~~ $\mathrm{O_2}$ bioflux is calculated with imperfect wind estimates from satellites or atmospheric reanalysis ~~.. The observed~~The observed$\Delta \mathrm{O_2/Ar}$ values are fixed and can not be adjusted to compensate for the errors in the winds used to calculate $\mathrm{O_2}$ bioflux. We test the effect of wind errors by using the reanalysis winds to drive the box model and QuikSCAT winds for the $\mathrm{O_2}$ bioflux calculation. The resulting $\mathrm{O_2}$ bioflux of such an experiment, shown as a green line in panel ~~D~~d, has a similar error (NRMSD~~=11.3%~~ $= 11.3\,\%$) to that of ~~in Panel C~~$\Delta \mathrm{O_2/Ar}$ in panel c. Finally, we compare the relative effect of imperfect wind estimates between ~~and~~ $\Delta \mathrm{O_2/Ar}$ and $\mathrm{O_2}$ bioflux. Time series of model NCP and wind from 500 locations in the Southern Ocean is used to force our boxmodel in an analogous fashion to the experiment presented in ~~figure 2, panel D.~~Fig. 2d in **?** . We find that $\mathrm{O_2}$ bioflux shows similar errors to $\Delta \mathrm{O_2/Ar}$, suggesting that both properties are useful for comparing the models and observations. We choose to use $\mathrm{O_2}$ bioflux since the units of flux are more appropriate for the current study.

~~This work was supported in part by funding from National Aeronautic and Space Administration (NASA NNX08AF12G) and National Science Foundation (NSF OPP-0823101).~~



~~Panel A shows mixed-layer integrated NCP from the TOPAZ model at 61.5°S, 159.5°W over a 400 day period. Panel B shows two wind time-series for the same time. The blue line is the NCEP/CORE reanalysis winds from the same location as the model NCP data, and the red line is Quickseat satellite-derived winds from the same location. The two lines in panel C represent the resulting O2/Ar supersaturation from a box model simulation based on the time series in panels A and B. Panel D shows O2 bioflux calculated from the time series in panels B and C. Red line is based on reanalysis winds and , blue line is based on Quickseat winds and , and green line is based on Quickseat winds and calculated from reanalysis winds.~~

Fig. 1. Map of O_2 observations used in the study. Grey shadings signifies the four regions we focus on in this study. Dots indicate discreet sampling locations while red lines indicated continuous sampling.

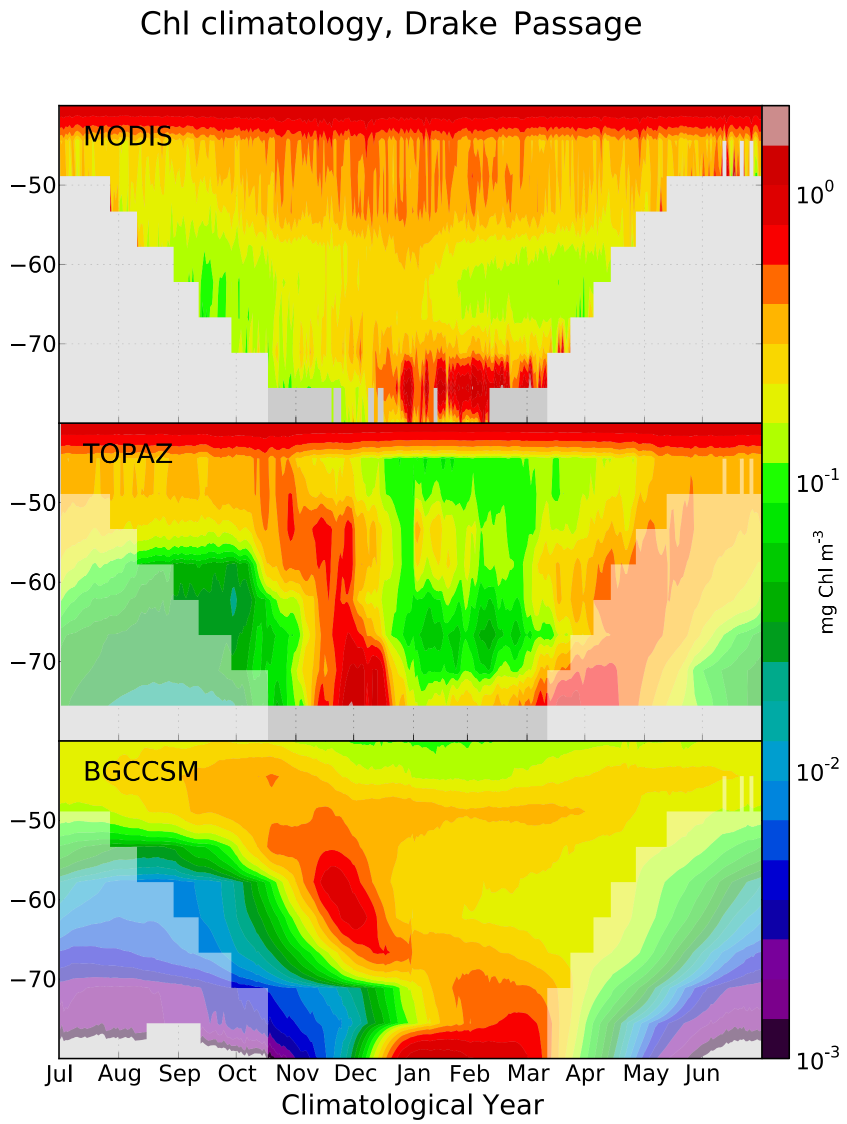


Fig. 2. Hofmøller plots of satellite Chl (top panel), Chl simulated by TOPAZ (middle panel), and Chl simulated by BGCCSM (bottom panel) in Drake Passage. All panels are zonal medians within the box. All data that fall on a specific grid cell on a specific year day are averaged to one value.

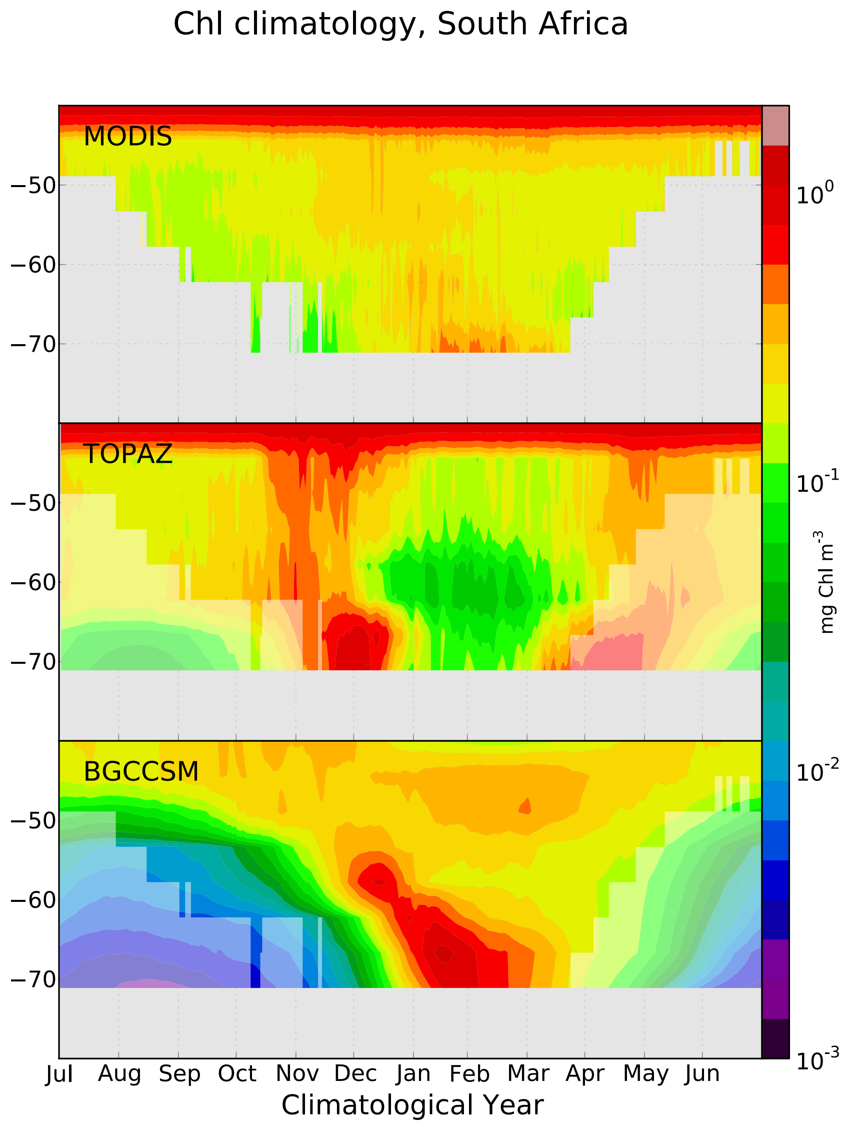


Fig. 3. Hofmøller plots of satellite Chl (top panel), Chl simulated by TOPAZ (middle panel), and Chl simulated by BGCCSM (bottom panel) south of South Africa. All panels are zonal medians within the box. All data that fall on a specific grid cell on a specific year day are averaged to one value.

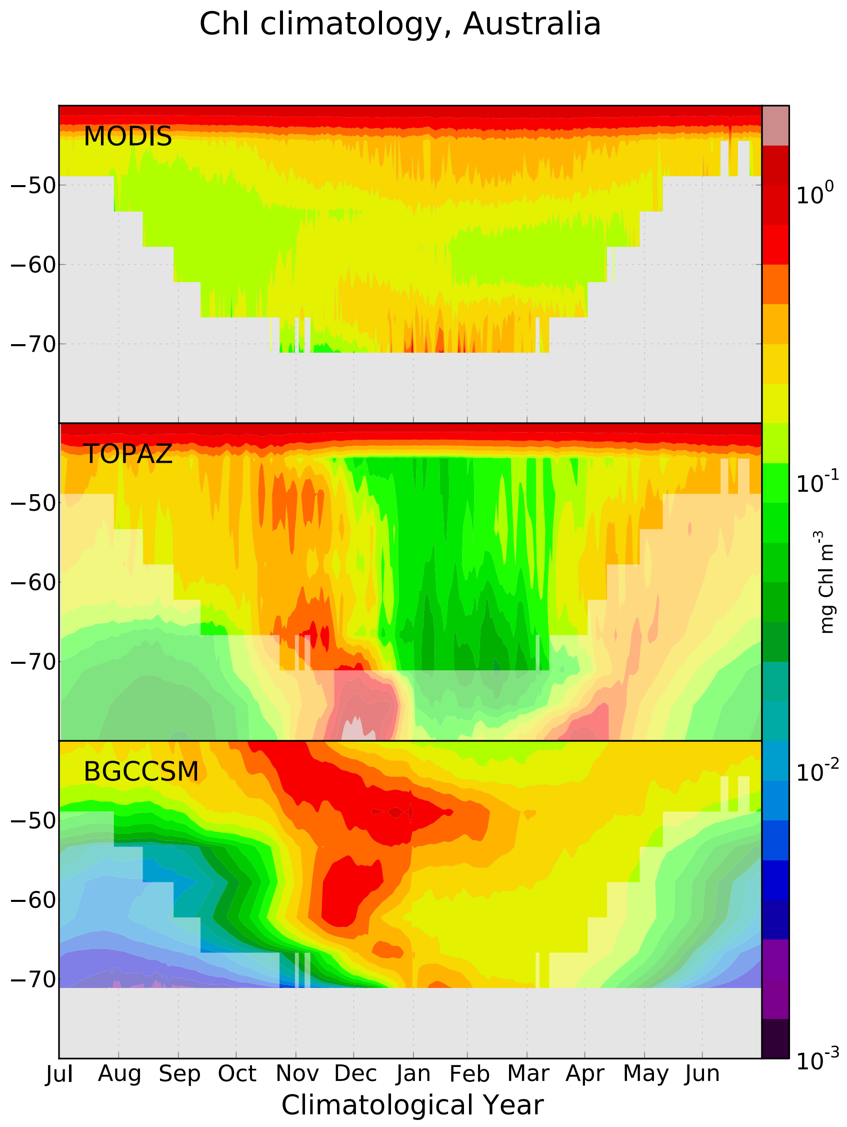


Fig. 4. Hofmøller plots of satellite Chl (top panel), Chl simulated by TOPAZ (middle panel), and Chl simulated by BGCCSM (bottom panel) south of Australia. All panels are zonal medians within the box. All data that fall on a specific grid cell on a specific year day are averaged to one value.

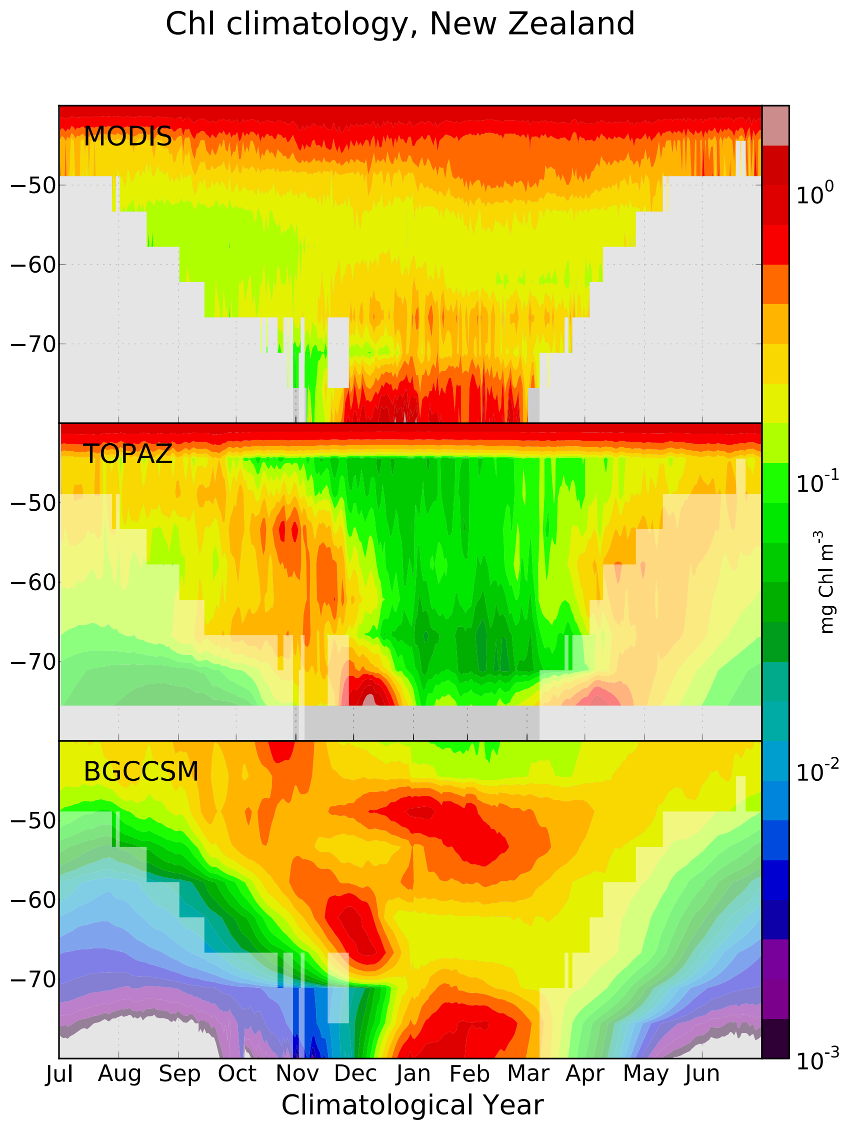


Fig. 5. Hofmøller plots of satellite Chl (top panel), Chl simulated by TOPAZ (middle panel), and Chl simulated by BGCCSM (bottom panel) south of New Zealand. All panels are zonal medians within the box. All data that fall on a specific grid cell on a specific year day are averaged to one value.

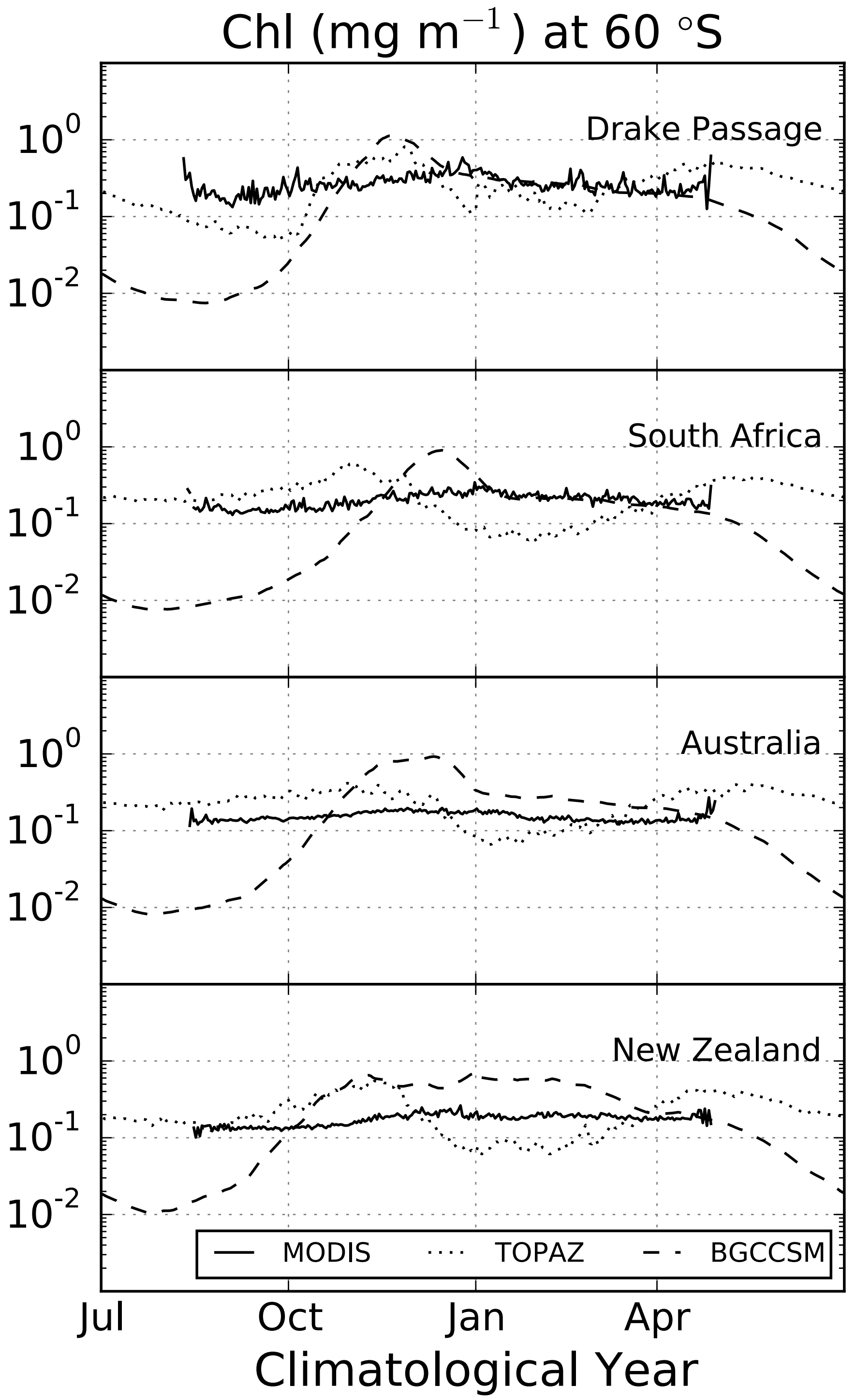


Fig. 6. ~~Hofmøller plots~~ Detail of ~~model NCP (top panel), model O2 bioflux (middle panel), and observed O2 bioflux (bottom panel)~~ the Chl concentrations in ~~Drake Passage (locations~~ mg m^{-3} from figures **??** - **??**. The lines show a a slice of ~~observations are presented as red points in the map)~~each original panel at 60° south. All ~~model values are zonal medians within the box. All observations~~ data that fall on a specific grid cell on a specific year day are averaged to one value.

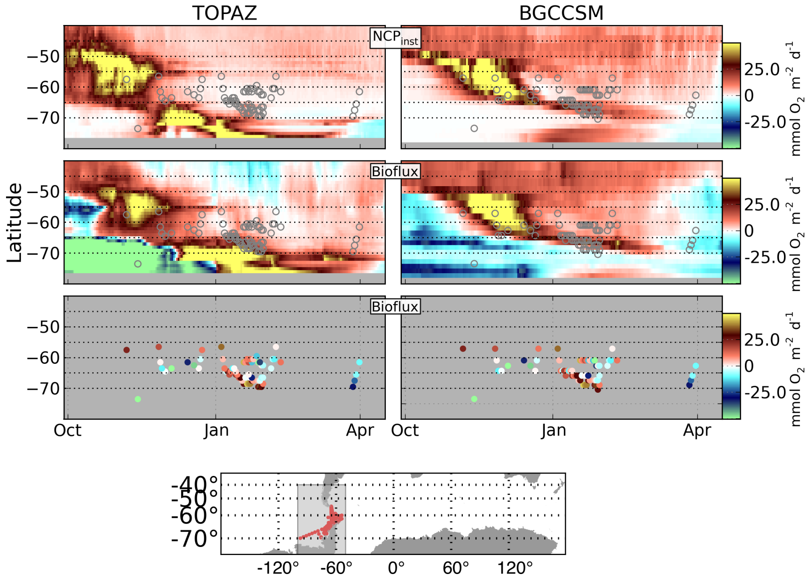


Fig. 7. Hofmøller plots of model NCP $(\mathrm{mmol\,m^{-2}\,d^{-1}}$, top panel), model O2 bioflux $(\mathrm{mmol\,m^{-2}\,d^{-1}}$, middle panel) , and observed O2 bioflux $(\mathrm{mmol\,m^{-2}\,d^{-1}}$, bottom panel) in Drake Passage (locations of observations are presented as red points in the map). All model values are zonal medians within the box. All observations that fall on a specific grid cell on a specific year day are averaged to one value.

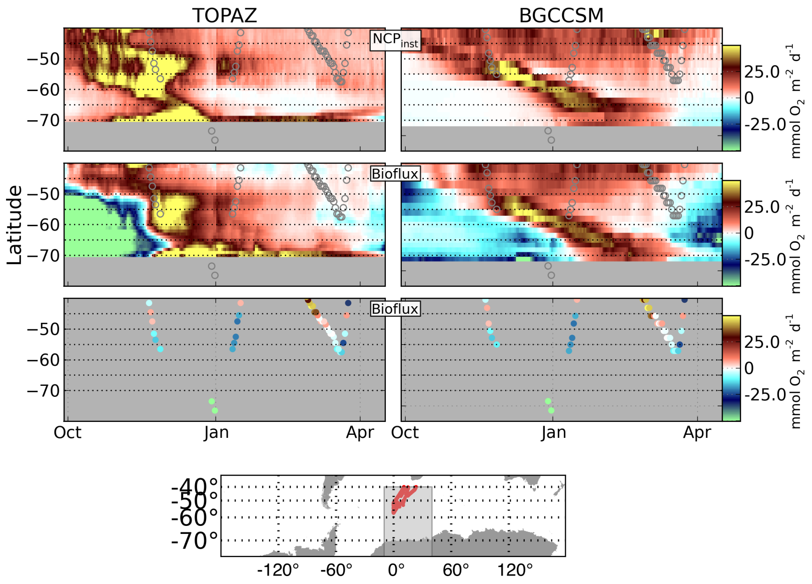


Fig. 8. Hofmøller plots of model NCP ($\mathrm{mmol\,m^{-2}\,d^{-1}}$, top panel), model O2 bioflux ($\mathrm{mmol\,m^{-2}\,d^{-1}}$, middle panel), and observed O2 bioflux ($\mathrm{mmol\,m^{-2}\,d^{-1}}$, bottom panel) in the Southern Ocean south of South Africa (locations of observations are presented as red points in the map). All model values are zonal medians within the box. All observations that fall on a specific grid cell on a specific year day are averaged to one value.

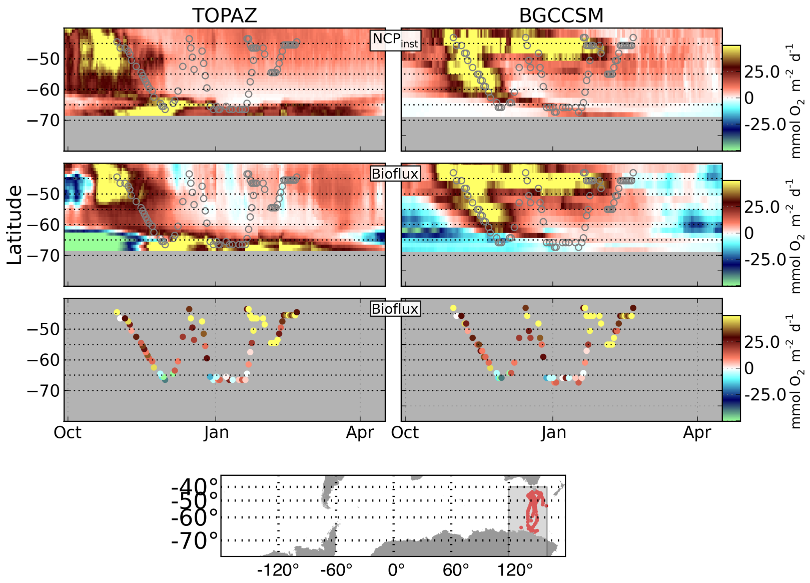


Fig. 9. Hofmøller plots of model NCP $(\mathrm{mmol\,m^{-2}\,d^{-1}}$, top panel), model O2 bioflux $(\mathrm{mmol\,m^{-2}\,d^{-1}}$, middle panel), and observed O2 bioflux $(\mathrm{mmol\,m^{-2}\,d^{-1}}$, bottom panel) in the Southern Ocean south of Australia (locations of observations are presented as red points in the map). All model values are zonal medians within the box. All observations that fall on a specific grid cell on a specific year day are averaged to one value.

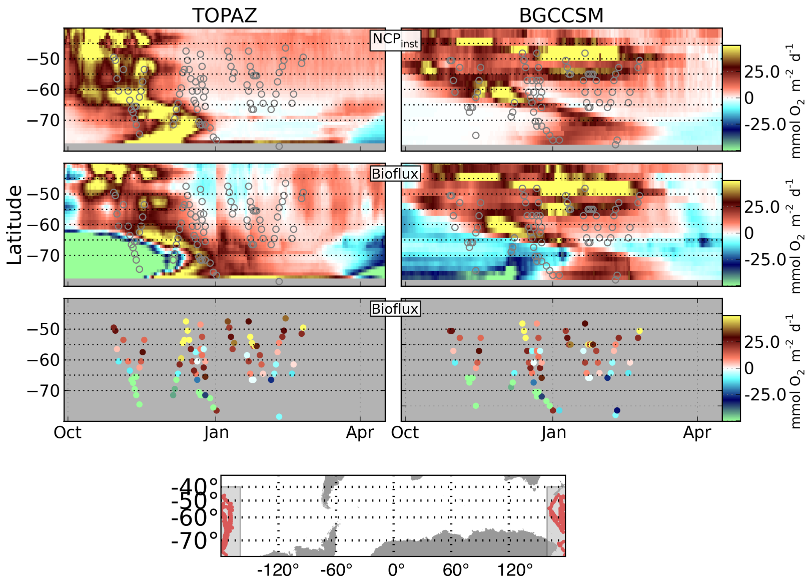


Fig. 10. Hofmøller plots of model NCP ($\mathrm{mmol\,m^{-2}\,d^{-1}}$, top panel), model O2 bioflux ($\mathrm{mmol\,m^{-2}\,d^{-1}}$, middle panel), and observed O2 bioflux ($\mathrm{mmol\,m^{-2}\,d^{-1}}$, bottom panel) in the Southern Ocean south of New Zealand (locations of observations are presented as red points in the map). All model values are zonal medians within the box. All observations that fall on a specific grid cell on a specific year day are averaged to one value.

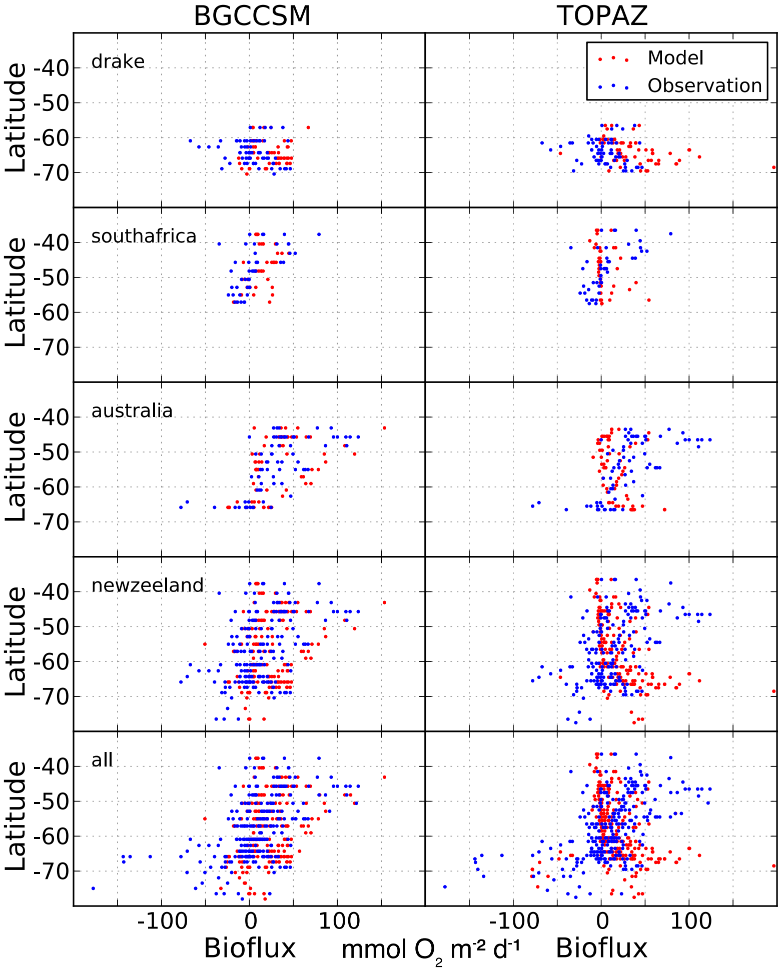


Fig. 11. Scatter plots of observed (blue) and model (red) O2 bioflux (mmol m^{-2} d^{-1}) versus latitude in four regions of the Southern Ocean. All observations that fall on a specific grid cell at a specific year day are averaged to one value. Models are subsampled at the location and year day of the observations.

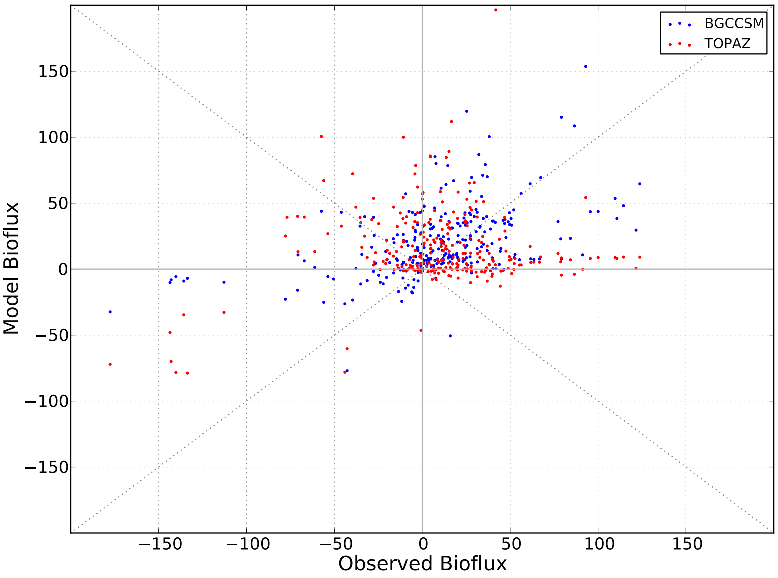


Fig. 12. Scatter plots of model versus observed O2 bioflux $(\mathrm{mmol\,m^{-2}\,d^{-1}})$ for the observational sampling sites shown in Figure B. All observations that fall on a specific grid cell at a specific year day are averaged to one value. Models are subsampled at the location and year day of the observations.

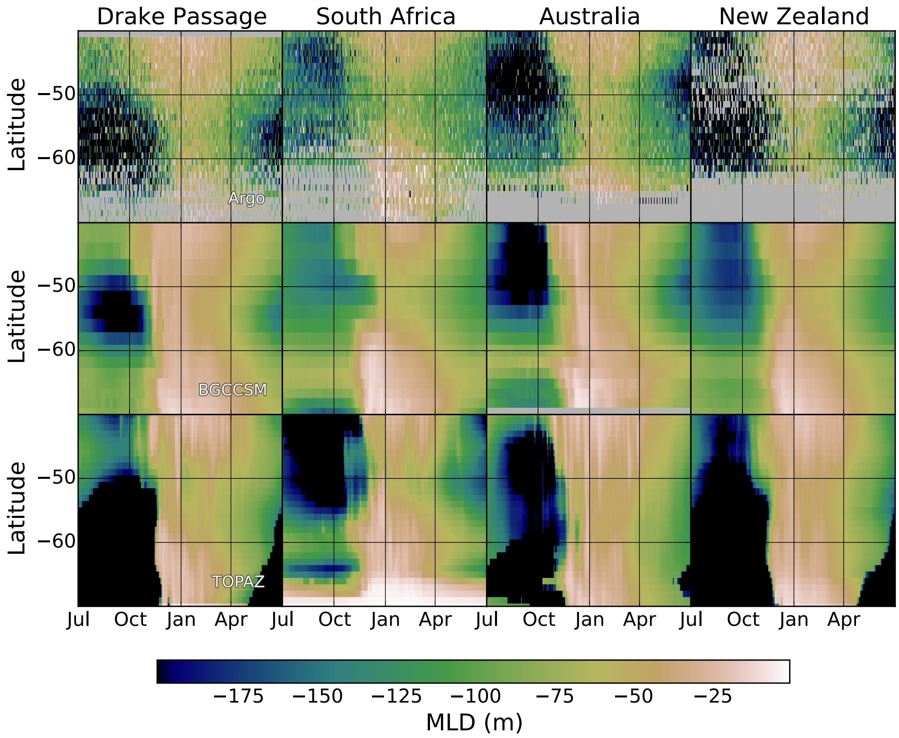


Fig. 13. Climatological Hofmøller plots of MLD (m in different regions of the Southern Ocean. All model values are zonal medians within the box. All observations that fall on a specific grid cell on a specific year day are averaged to one value.

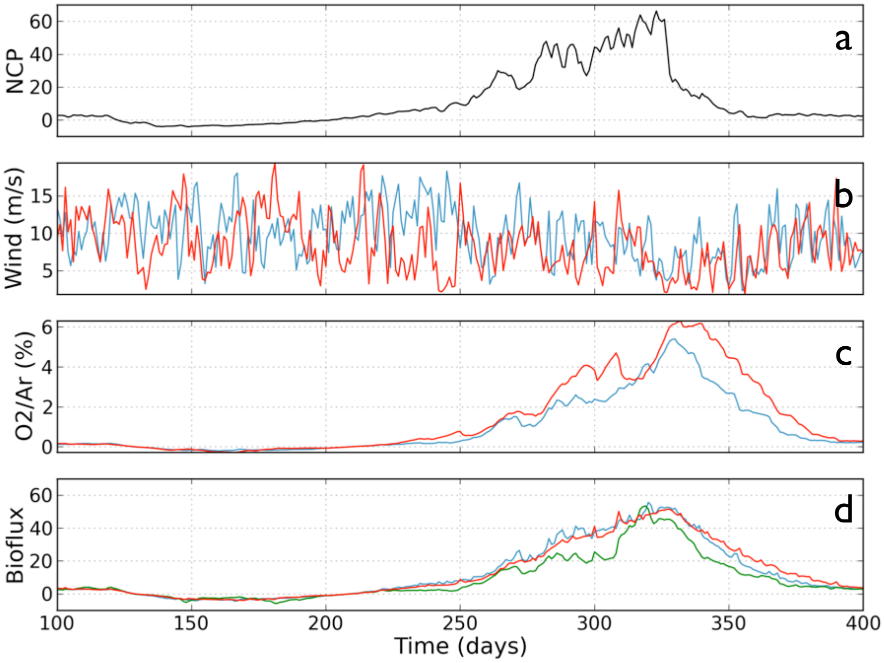


Fig. 14. Panel A shows mixed-layer integrated NCP from the TOPAZ model at 61.5°S, 159.5°W over a 400 day period. Panel B shows two wind time-series for the same time. The blue line is the NCEP/CORE reanalysis winds from the same location as the model NCP data, and the red line is Quickscat satellite-derived winds from the same location. The two lines in panel C represent the resulting $\Delta O_2/Ar$ supersaturation from a box model simulation based on the time series in panels A and B. Panel D shows O2 bioflux calculated from the time series in panels B and C. Red line is based on reanalysis winds and $\Delta O_2/Ar$, blue line is based on Quickscat winds and $\Delta O_2/Ar$, and green line is based on Quickscat winds and $\Delta O_2/Ar$ calculated from reanalysis winds.