Manuscript # bg-2015-296: Response to referees

We thank the three referees for taking time to review our manuscript "***Marine regime shifts in ocean biogeochemical models: a case study in the Gulf of Alaska***" and for providing thoughtful and thorough comments. We agree with the referees' main suggestions, and have revised the manuscript to address those points.

Referees comments are in bold and our responses in plain text.

Response to anonymous referee 2:

I think this is a good paper that is publishable with relatively minor revisions, assuming that none of the things I flagged as insufficiently explained lead to discovery of major conceptual issues. I congratulate the authors on a generally well written paper.

We thank the referee for this positive view of our manuscript and for his/her constructive suggestions, which significantly helped improve our manuscript.

(1) A major conceptual issue is that the authors treat 'regime shifts' as being qualitatively different and distinct from 'red noise' (e.g., 14015/9-11, 14022/13-14), and I don't think there is a strong conceptual basis for this. North Pacific variability follows a red noise spectrum, and within such a spectrum will inevitably be found some brief periods of rapid change that can be interpreted as regime shifts. This is the key point of Rudnick and Davis. The point is precisely that regime shifts are part of the continuum of variability, not a priori evidence of shifts between discrete stable states, as this paper seems to imply. The assertion that a "change in the slopes rather suggests a change in the relationship and thus, a nonlinear response" (14016/2-3) directly contradicts Rudnick and Davis who state that detection of regime shifts is "not evidence of nonlinear processes leading to bi-stable behavior".

This is an excellent point. For the purpose of this manuscript, we define a shift in a statistical sense, i.e. a shift in the mean or in the parameters of a regression model, which is unusual given the background autocorrelation and length of the time series. This is already an improvement over several regime shift detection methods not making this distinction – and interpreting trends as regime shifts. The problem of ultimately deciding whether a time series is better fitted with a red noise model (AR(1) or even longer and more complex memory) as opposed to a change-point model representing several stable states, is very complex as these different effects interact with estimation of parameters . As far as we are aware of, and after consulting with other statisticians' colleagues, there is no objective criterion to decide whether a time series is better fitted with an AR(1) rather than a change point model. This is actually the subject of an ongoing project, which is beyond the scope of this manuscript.

Given the occurrence of one significant shift in the available record, we feel that the underlying assumptions of a change point with background AR(1) are appropriate. However, this assumption may not be appropriate at a finer time scale (e.g. monthly). In response to this valuable comment, we now express this caveat in the discussion and have adjusted the language through the manuscript accordingly.

(2) I also think the discussion of 'predicting' regime shifts with coupled models (top p. 14023) is vague and overoptimistic. Simulating such events with a forced ocean model (hindcast) and with a coupled model are very different propositions. Predicting them is much more difficult still. No results shown in this paper have any bearing on whether such predictability is possible. It may be that these authors are simply misusing the word 'predict' and don't actually mean this at all (the final sentence of this paragraph, discussing downscaling of climate projections, suggests that this is the case). But in any case I don't think this section is useful; it could be substantially reworded or deleted entirely. All that has been demonstrated here is that the ocean model is adequate to simulate the PDO mode in a hindcast, and therefore it is *possible* that the mode could be accurately simulated (in a statistical sense) in a coupled model. Projections with such a model could be usefully downscaled, but this does not in any meaningful sense constitute a 'prediction' of future regime shifts. I recommend the authors go through the MS searching on every instance of 'predict' or 'prediction' and consider carefully whether it is (a) accurate and (b) necessary. (I would do the same with "nonlinear" in accordance with point (1) above.)

We agree with the referee that the prediction of regime shifts with coupled models is overoptimistic and vague. To avoid any confusion, we decided to entirely delete this paragraph of the discussion section. We also adjusted the language when referring to nonlinear changes.

(3) There are several implausible elements in the data shown in the graphics. In Figure 5 the black dashed lines are said to represent regression equations over the whole half century from 1957-2007. But these lines do not seem very plausible to me. In the case of chlorophyll, if we took this line and rotated it about 20-30 degrees clockwise it would be a much better fit. The residuals would be smaller and much more homogeneous. So it's hard to envision a procedure that would generate this line as the least-squares best fit to these data. This applies to the other panels and Figure 6 as well, although for 5b (PP) it's a bit hard to tell because there really do seem to be distinct regimes before and after 1977 and any single relationship would be a poor fit.

Similarly I find it implausible that the zooplankton time series in Figure A1d is best fit by a constant value rather than a regime shift in 1979-1980 as is the case for CHL, PHY and DIN in this model. Whether or not the pre-1979 values are best fit with a constant or a slight downward slope, I find it very hard to believe that the least-squares fit would not improve if the post-1980 mean was reduced by about 0.01. Or perhaps a single long term downward trend (Model III) would be the best fit. But it isn't plausible that the model shown in the figure is the best one. This suggests that the method applied is not quite as general as presented and that some unstated assumptions may have been inadvertently coded into the statistical procedure.

We appreciate this comment about the statistical fit, to which we subsequently repeated and verified all analysis to ensure that our results are robust and accurate. We do not find that any of the elements shown in the graphics are implausible. For example, when looking at Figure 5 a), the referee suggestion to rotate the overall fit (black dotted line) by 20 degrees seems influenced by one single large chlorophyll value (the one point >1.1 mg/m3), which would not be an objective fit in the least squares sense.

As for Figure A1, the best fit in panel d) is indeed one with a shift in 1977 and the straight line is just a plotting mistake. Table 3 does show that a shift is more likely in the HadOCC ZOO time series. We verified that all fit plotted are consistent with the results presented in Table 3 and found no other problems. The fits selected represent a compromise between the likelihood of the fit penalized by the number of parameters, that we then further test to assess significance if a model with a shift is selected. Some of these time series contain autocorrelation, which is taking into account in the Monte Carlo simulations, but to make sure it is transparent to the reader that we actually do, we also specify the time series containing autocorrelation in Table 3.

(4) The principal components analysis is not explained in the Methods. The caption to Figure 4 implies that the PC's were calculated for regional averages, but the region of averaging is not stated. Such averaging is not necessary to calculate principal components: one could just combine all of the variables into one big state space and calculate EOFs for that space. But either way you need to specify the geographic domain, or the region of averaging if regional means were used. It also usual to normalize the different fields (e.g., z-scores) so that the different magnitudes of the variables (and therefore arbitrary choices of units) do not affect the results. I assume this was done but it is not stated. I also assume they used annual rather than monthly data, but again this is not specified in the text. There should be a short paragraph in the Methods describing exactly what was done here.

The time and spatial averaging is explained in section 2.2: "The time series were averaged from monthly means to annual means and then averaged spatially across the region defined by the boundaries of 54°N to 62°N and 130°W to 160°W (same region as the observational dataset used, see section below)." These spatially averaged time series were used in the principal component analysis as well, and this is now mentioned in the methodology section. After double-checking the pca analysis, we realized that it was not performed on the z-scores, but agree with the referee it should have been. Therefore, the results in tables 4 and 5 have been updated and Fig. 4 was replaced (see Fig. 4 below). We also removed analysis performed in PC2 from the manuscript for simplicity, as it did not show anything worthy to discuss in the manuscript. The values presented in the tables slightly vary, but do not affect any of the conclusions. We really appreciate that the referee flagged this, which improved accuracy of our manuscript. As suggested by the referee, we added more details into the methodology section so the reader can follow more easily what we have done.

(5) I think the biogeochemical model descriptions could be clearer, particularly in the area of phytoplankton nutrient limitation. Some models are described as using multiplicative limitation and others as employing the "law of the minimum", but the description is vague with respect to which environmental factors these formulations apply to (e.g., 14010/24-25, 14011/12-13, 24). The usual practice is to use a minimum for multiple nutrient limitations (Blackman's rule), and then either a multiplicative or a minimum for nutrient vs light limitation (and occasionally temperature also although models that use a min()

function for temperature are rare). I don't know of any model that uses a multiplicative function for e.g. N and P limitations.

We thank the referee for this comment. We added some clarifications about phytoplankton nutrient limitation in the biogeochemical model description in section 2.1. In summary, HadOCC, Diat-HadOCC and MEDUSA all use multiplicative nutrient limitation, where light limitation is multiplied by successive nutrient limitation terms. ERSEM uses a combination of multiplicative and maximum limitation factors. PlankTOM uses the minimum of nutrient limitation terms to regulate phytoplankton growth.

(6) The data presentation is uninspired. Figures 5-8 all show variations on the same thing, and show only a limited and arbitrary subset of the possibly configurations (4 biogeochemical fields vs SST for two models and vs MLD for two others). These plots are somewhat space-inefficient, and in principal these authors could show many more data in fewer figures e.g., by showing a 4x5 matrix of CHL/PP/PHY/ZOO (rows) vs all five models (columns) for SST (Figure 5) and MLD (Figure 6). Maybe they don't think it is necessary to show all of the data, but right now it seems like only an arbitrary subset are shown. Even if only the current set are shown, the four figures could be reduced to two as there is a lot of whitespace and redundant information.

We concur. We have merged Figs. 5-6 together in the revised version, which is much more space efficient. We show a subset of the data as we conducted this analysis only for the models that exhibit a significant shift in the late 1970s in one of the principal components, this is specified in the Methods and Results sections.

The footnotes to Table 4 are confusing and unnecessary. c and d do not appear to be used at all. a and b are not necessary as they are redundant to information already in the Table. All that is needed is to add "Years in bold have a significant shift" to the caption. I am generally opposed to the practice of specifying significance levels as P=X rather than P<alpha, but in this table both are used. Choose one or the other. This applies to Table 6 as well, except in this case note c does appear a few times.

As suggested by the referee, we have removed the signs indicating different significance level in Table 4 footer and indicate significance by bold fonts. To have a consistent style, similar changes were applied to tables 3, 5 and 6. To keep the presentation of the tables tidy, we present the p-values as p=X with two decimal places. The cases for which the p-value is very small are noted as p<0.01.

I don't think equations 2, 3, 5 or 6 are necessary. What information do they contain that is not already expressed in equations 1 and 4?

We have removed Eqs. 2 and 3 and rather just introduce the notation, i.e. "The most likely timing for a shift under models IV and V can be found similarly, and are noted $SIC_{IV}(p)$ and $SIC_{V}(p)$, respectively.“ However, we believe removing Eqs. 5 and 6 might lead to confusion from the readers as the null hypotheses in these two decision rules is different than the one in equation 4, thus we kept these two equations.

Table 2 add space after epsilon in first line

Space added.

Multiple references within a parenthesis should ordered either alphabetically or chronologically. I don’t know if this journal specifies which but it should be one or the other.

According to the style of the journal, they should be ordered chronologically. We carefully checked the whole document to make sure the referencing is consistent and corrected the cases that were not ordered chronologically.

14011/19-20 I’m not sure what is meant by ’heterotrophs’ here (“three zooplankton
groups (heterotrophs, microzooplankton and mesozooplankton)”). Aren’t all zooplankton
heterotrophic?

Yes, we meant heterotrophic flagellates and corrected it in the new version of the manuscript.

14005/9 delete “itself”
14006/9 delete “number”
14007/3 change “distinguish” to “distinguishing”
14007/24 delete “Specifically”
14008/14 change “challenged” to “limited”
14008/29 change “described” to “ascribed”
14010/8 phosphorus misspelled
14014/22 Not clear what “length” means in this context.
14015/19 change “support” to “aid”
14018/4 add reference to Table 5 here; add “principal” before “component”

All these edits were made as suggested by the referee.

14007/24-26 The interpretation of Polovina et al 1995 here strikes me as overly simplistic. If you look at their Figure 8C, whether mixed layer depth and zooplankton biomass increase or decrease depends on the season. I think it’s fair to say that the MLD shoaled after 1977 (their Figures 3 and 5C), at least in winter, and that the winter mixed layer depth probably drives the seasonal cycle of biological productivity. But what is stated here is not an accurate characterization of the data shown in that paper, and anyway the model used is rather archaic and maybe shouldn’t be taken too seriously.

We thank the referee for this comment and decided to shorten the discussion about the results of Polovina et al. (1995) and instead we added further discussion about a few more recent modelling studies.

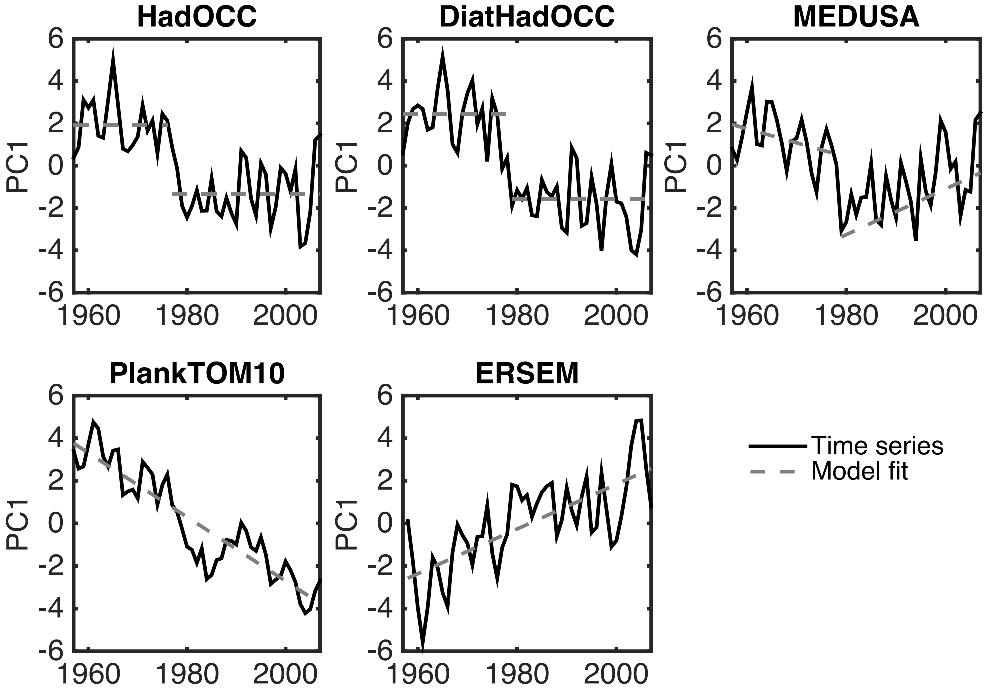

Figure 4. First principal component (PC1) of sea surface temperature, mixed layer depth, surface dissolved inorganic nitrogen, silica, iron, surface chlorophyll, integrated primary production, total surface phytoplankton and zooplankton biomass (if available) averaged over the Gulf of Alaska for each model for each model.

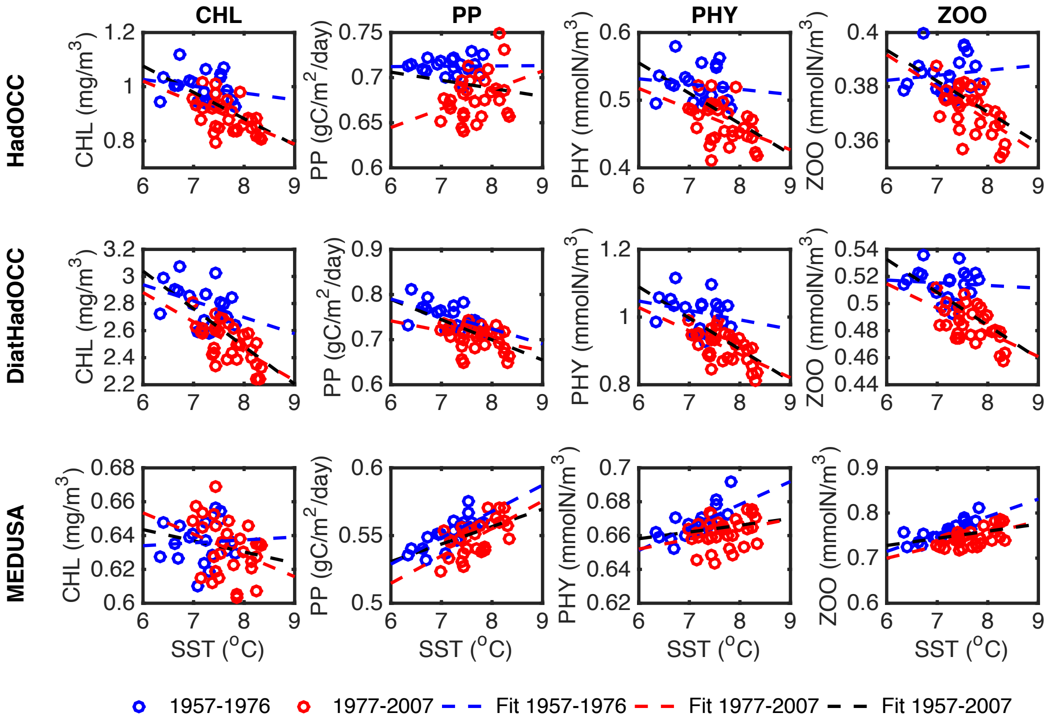

Figure 6. Relationships matrix between simulated sea surface temperature (SST) and the biological variables over the Gulf of Alaska region. Rows represent different models (HadOCC, DiatHadOCC and MEDUSA) and columns represent different biological variables (surface chlorophyll (CHL), integrated primary production (PP), total surface phytoplankton (PHY) and zooplankton biomass (ZOO)). Linear relationships are inferred for the periods 1957-1976, 1977-2007 and 1957-2007 using least square regression. Table 5 presents test results on the similarity of these relationships.