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, which has substantially improved the manuscript. We agree with the referees' main suggestions, and have carefully revised the manuscript to address those points.

In summary, here is the list of all relevant changes:

1) Additional paragraphs in the introduction/discussion, such as additional comparison to previous modelling studies in the region, effects of model complexity and methodology limitations.
2) Discussion and additional analysis on the PDO, ENSO and NPGO time series– added in Figure 3.
3) Updated principal component analysis focusing on PC1 only, as well as additional details in the methodology section. Previous results on PC2 were removed from Tables 3-5 and from Figure 4. These changes also led to move a table (Table 3 in the previous version) to the Appendix.
4) The forcing-response analysis is done on the three models that exhibit a shift in PC1. Figs. 5-8 were combined into one and Table 5 updated accordingly.
5) A new figure presenting timing of shifts in gridded SST was added, unveiling coherent spatial patterns of late 1970s shift in and around the Gulf of Alaska (Figure 2).

Below, we present the detailed responses to the referees' comments. In our responses, we point out where the change was made and include modified figures where appropriate.

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

Response to anonymous referee 1:

This paper explores the response of five different ocean biogeochemical models averaged over the Gulf of Alaska (GoA) to the same physical forcing over the period 1950-2007. Time series of physical (SST and MLD) chemical (DIN, SI, FE) and biological (surface chlorophyll, integrated primary productivity, and surface phytoplankton and zooplankton biomass) from the models are analyzed using a change point detection scheme. The method consists of a set of regression models that can classify the time series as a constant mean, mean shift, trend, shift in the intercept of the trend and shift in both the intercept and trend. The method is able to detect one change throughout the time series. A downward trend in GoA SST is identified prior to 1976 followed by a weak upward trend afterwards with a slight upward trend in MLD over the period. Most of the simulated biogeochemistry time series indicate a change around 1976 but show a mix of behavior with the simpler models exhibiting more regime-like behavior than the more complex models. A comparison of how different ocean biogeochemical models with the same physical state simulate time series of key quantities in different parts of the globe, including the Gulf of Alaska, is a useful endeavor. The same can be said for a fairly rigorous evaluation of change points in these time series. Thus I accept this paper for publication in Biogeosciences, but I think the manuscript can be improved in several ways and thus I recommend a major revision.

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

1) While regime analysis has become very popular especially in climate and marine ecosystem analysis, it may lead to a misinterpretation of the underlying dynamics of a system, especially for relatively short time series. For example, regimes are often linked to phases of the Pacific Decadal Oscillation (PDO). The transition in the North Pacific around 1976-77 has been linked to a change in the PDO (which extends into the tropics), while the one around 1998 is more associated with the second EOF sometimes termed the North Pacific Gyre Oscillation (NPGO, also relevant for the discussion in the Introduction, page 5, lines 22-33). Rather than regimes these just might be periods where one pattern is more prevalent than another, where both of these patterns impact the GoA. In addition, evaluating the time series as single AR1 process (for the model of the mean and no change) may not be the best null hypothesis. Anomalies in the state of the GoA are strongly influenced by ENSO and other factors. If these processes fluctuate, they could cause rapid changes in the remote time series even if these processes are linear and add to each other. The authors should discuss these complicating factors (or perhaps even try and incorporate them as one of their models. See the following papers:
Bond, N. A., J. E. Overland, M. Spillane, and P. Stabeno (2003), Recent shifts in the state of the North Pacific, Geophys. Res. Lett., 30(23), 2183, doi:10.1029/2003GL018597. B
Di Lorenzo, E. et al. North Pacific Gyre Oscillation links ocean climate and ecosystem change. Geophys. Res. Lett. 35, L08607 (2008).

Matthew Newman, 2007: Interannual to Decadal Predictability of Tropical and North Pacific Sea Surface Temperatures. J. Climate, 20, 2333–2356. doi: http://dx.doi.org/10.1175/JCLI4165.1
Newman, M., G. P. Compo, M. A. Alexander, 2003: ENSO-forced variability of the Pacific Decadal Oscillation. J. Climate, 16, 3853-3857.
Schneider, N., and B. D. Cornuelle (2005), The forcing of the Pacific Decadal Oscillation, J. Clim., 18, 4355 – 4373.

We highly appreciate this suggestion. We added a paragraph in the discussion and conclusions section (second paragraph) where we discuss potential contribution of the PDO, NPGO and ENSO to the underlying dynamics based on the papers suggested above. If both the PDO and NPGO fluctuations drive changes in the North Pacific climate and ecosystem functioning (and implicitly ENSO which influences the PDO), the question arises whether either of these indices exhibit a shift at a similar time. We have included these three large-scale oscillations in our analysis and verify whether they also exhibit a change-point in the late 1970s (Fig. 3, presented at the end of our response). As mentioned in previous work, the PDO index exhibits a significant shift in 1976/77, but we find no significant shifts in the multivariate ENSO index or the NPGO index. Clearly, by detecting a shift in the late 1970s in PDO only we cannot conclusively tie the PDO and untie the NPGO and ENSO to the shift in climate and ecosystem dynamics of the Gulf of Alaska, but this provides useful information for future work aiming at determining causal mechanisms, teleconnections and physical/biogeochemical dynamics linking global climate patters to ocean productivity in this region.

As for using an AR(1) as a null model, we think that it is an appropriate null hypothesis at the annual time scale and should act as a parameter which roughly comprises these complex factors. The need to add an autoregressive model to a statistical model usually indicates some external factors have not been taken into account in the fit. Here, we summarized these external factors through an AR(1), as our objective is not to attribute external factors influencing physical/biogeochemical variables in the Gulf of Alaska, but rather assess whether they exhibit significant abrupt changes. This point is now clarified in the methodology section.

2) How would the method used here classify a pure sine wave with (one) zero value some where in the time series? Would it classify the zero crossing as a change point or regime shift? (Same goes for the changes in the "trends" when the amplitude of the waves switches sign.) Clearly the dynamics behind an oscillatory signal would likely be quite different than the dynamics for a regime shift or trend.

Yes, the dynamics behind an oscillatory signal would be different than the dynamics for a regime shift or trend. However, given that the models take into account the autocorrelation, we think the oscillatory signal would be interpreted as temporal correlation, which the autocorrelation component should pick up.

3) The authors should probably reference other modeling studies of decadal physical and biogeochemical changes in the northeast Pacific:

Alexander, M., A. Capotondi, A. Miller, F. Chai, R. Brodeur and C. Deser, 2008: Decadal variability in the Northeast Pacific in a physical-ecosystem model: The role of mixed layer depth and trophic interactions. Journal of Geophysical Research - Oceans, 113, C02017, doi:10.1029/2007JC004359.
Capotondi, A., M. A. Alexander, C. Deser, and A. J. Miller (2005), Low frequency pycnocline variability in the northeast Pacific, J. Phys. Oceanogr., 35, 1403 – 1420
Haigh, S. P., K. L. Denman, and W. W. Hsieh (2001), Simulation of the planktonic ecosystem response to pre- and post-1976 forcing in an isopycnic model of the North Pacific, Can. J. Fish. Aquat. Sci., 58, 703 – 722.

We thank the referee for bringing these studies to our attention. We have included a new paragraph in the discussion and conclusions section (third paragraph) in which we compare our results to these modelling studies. We feel this suggestion really improved our discussion. Our results are in agreement with Haigh et al. (2001), who are suggesting a year-round deepening of the mixed layer depth in the Gulf of Alaska after 1976, which led to a decrease in nutrient, phytoplankton and zooplankton after 1976. Alexander et al. (2008) rather simulate a shoaling in the winter mixed layer depth in the late 1970s, giving rise to a early-spring increase in primary production, phytoplankton and zooplankton biomass followed by a late-spring decline in both phytoplankton and zooplankton biomass. Despite the caveat that we are analysing annual mean time series it is important to point out the contradictory direction of change in mixed layer depth. Possibly reconciling this discrepancy, Capotondi et al. (2005) suggest a deepening trend in a broad band along the coast and shoaling in the central part of the Gulf of Alaska. Thus, the comparison of the various attempts to simulate the late 1970s regime shift of the Gulf of Alaska raises the possibility that the abrupt and spatially coherent ecosystem change is actually caused by a previously unappreciated heterogeneous set of environmental changes with distinct spatial pattern and timing in the annual cycle. Further analysis would be required to investigate changes at the seasonal scale and at a finer spatial resolution and is beyond the scope of this study.

4) The reference for Wunsch 1999 on page 4 line 20 is missing from the reference list.

Reference added – we thank the referee for pointing this out.

5) The authors should probably include a discussion of the physical model simulation here even if it is described in other papers and/or on line. Are the BGC models driven offline where the ocean model (NEMO) is run first and then the values are fed into the BGC models? Note this does not allow for feedback of the biology on the physics (e.g. changes in solar absorption by phytoplankton). If the surface sensible and latent heat flux are computed using the observed air temperature and model SST, the results will be to strongly relax the model SST towards the observed SST (as the observed air temperature & SST are highly correlated. If this is the case then the model will likely obtain the correct SST regardless of if it is a good simulation or not, e.g. see: Seager, R., Y. Kushnir, and M. A. Cane (1995), On heat flux boundary conditions for ocean models, J. Phys. Oceanogr., 25, 3219 – 3230.

We thank the reviewer for this suggestion that clarifies the simulations that were used. The simulations were "online", in that physics and biogeochemistry were both formally simulated simultaneously. Feedbacks between the model biology and ocean physics (e.g. by the absorption of downwelling solar radiation) were disabled so that all of the biogeochemical models experienced consistent simulated physics. In addition, sea surface salinity was weakly relaxed (characteristic timescale of 180 days) towards observations to minimise drift, while model SST was not relaxed to observed SST. These details have been added to section 2.2.

5) During late winter in the subarctic North Pacific the mixed layer extends to the upper portion of the halocline, located between depths of approximately 70 and 120 m (Roden, 1964; Freeland et al., 1997; de Boyer Montegut et al., 2004) and the MLD is mainly controlled by salinity not by temperature (this would have impacted the Polovina et al. [1995] finding). Low-frequency changes in the Ekman pumping in the Gulf of Alaska, which vertically displaces the halocline, may impact the wintertime MLD by moving a layer with strong density gradients toward or away from the surface. After the mid-1970s the pycnocline was shallower in the central part of the Gulf of Alaska and deeper in a broad band along the coast, primarily due to the local response to Ekman pumping (Cummins and Lagerloef, 2002; Capotondi et al., 2005). This impact should be included using a density definition for MLD although a change in MLD about 1976-77 seen in other studies is not found here.

de Boyer Montegut, C., G. Madec, A. S. Fischer, A. Lazar, and D. Iudicone (2004), Mixed layer depth over the global ocean: An examination of profile data and a profilebased climatology, J. Geophys. Res., 109, C12003, doi:10.1029/2004JC002378.
Capotondi, A., M. A. Alexander, C. Deser, and A. J. Miller (2005), Low frequency pycnocline variability in the northeast Pacific, J. Phys. Oceanogr., 35, 1403 – 1420.
Cummins, P. F., and G. S. Lagerloef (2002), Low frequency pycnocline depth variability at station P in the northeast Pacific, J. Phys. Oceanogr., 32, 3207 – 3215.
Freeland, H., K. Denman, C. S. Wong, F. Whitney, and R. Jacques (1997), Evidence of change in the winter mixed layer in the northeast Pacific Ocean, Deep Sea Res., Part I, 44, 2117 – 2129.
Roden, G. I. (1964), Shallow temperature inversions in the Pacific Ocean, J. Geophys. Res., 69, 2899 – 2914.

We indeed used a density dependent mixed layer depth, which should take account of these effects. However, we cannot separate central and coast MLD as we use a basin-wide mean time series of MLD (shallower in the middle and deeper along the coast). We added a section in the discussion and conclusion (third paragraph) to highlight this difference. Please also see response to comment 3.

7) Page 15. Paragraph 12-20. The authors indicate that several of the models depict a regime shift in the Gulf of Alaska in late 1980s (instead of the mid 1970s) and that this shift seems to be mainly forced by changes in MLD. However, the change detection method and Figs. 2-4 do not appear to show much of a change in MLD

around 1989 either in the full time series or in the PCs.

We agree and have now removed discussion about a late 1980s shift as this is not detected in the individual time series neither in the PC1 of the five OBGC models. We feel the manuscript is more focused now.

8) Give the correlation values for the curves in Fig. 7 & 8. Are the correlation values during the different epochs significantly different from each other (rather than significant – i.e. different than zero). The lines in Figs. 7 do not seem significantly different from each other, especially given the large spread (see 7 above).

In the revised version of the manuscript, previous Figs. 7-8 have been combined into one figure for less repetition and more efficient use of space (Fig. 6, presented below). The corresponding slope values are presented in Table 5, which we now also mention in the figure caption for clarity. We are actually testing whether the relationships before and after the shift are different from each other, not whether they are different from zero. They are mostly not significantly different (see Table 5).

9) Bottom of page 16 top of 17 (also in the abstract). The authors indicate that all models simulate a decrease in nutrients and biological productivity after 1976. Perhaps, this statement is based on Fig. 3; however, an examination of Figs. A1-A5 indicates more complex behavior. For example, the change point analysis suggests a downward trend for PHY & ZOO for the DiatHadOCC and PlankTOM10 models and an upward trend in FE in the ERSEM model over the entire record.

We reworded this part to reflect that this statement is true only for the models simulating the shift: "A shift in model SST occurred in 1976 and matched a shift in observed SST. This abrupt change was accompanied by a smooth deepening of the mixed layer depth followed by an overall decrease in nutrients and productivity. The three OBGC models simulating an abrupt change in 1977 in PC1 (i.e. HadOCC, DiatHadOCC and MEDUSA) are consistent in the overall direction of change (Fig. 4)."

10) Bottom of page 17 top of 18. A point of clarification . . . "20th century" simulations from the CMIP5 archive (the simulations that are referenced the most from the archive) do not produce a climate shift in the mid 1970s. These models are initialized in the mid 19th century and due to chaotic interactions values during a given time in the model do not directly correspond to those in nature (although the idea is that the models have the correct sensitivity to climate change and have the basic statistics of climate variability correct.) The Meehl and Teng studies (including the one referenced here) are based on initialized hindcast model runs just within a few years (up to 10) prior to the period examined.

We agree that this part of the discussion was overoptimistic and vague and decided to remove this paragraph from the discussion section in the revised version of the manuscript.

11) The authors note that the simpler models tend to produce more regime-like behavior. Are there references from other fields, e.g. systems theory, which can support this from a more general perspective?

See response to comment 12 below.

12) While the authors note that the models produce different change points, they don't comment much on the difference between models. Indeed one is struck by how different the simulations are especially given that the physical forcing is identical. What does this say about the state of ocean BGC modeling? Are there observations say at OWS P that could support one model over another? Are the BGC models highly nonlinear in that one should perform multiple ensembles (based on different initial BGC conditions) as is done for climate system forecasts – i.e. one would get different results from individual ensemble members. If the model results are so different, does that suggest caution in using change point analyses?

We thank the reviewer for this comment, which helped improve the discussion. The performance of the models in terms of a fit to observations has been assessed globally and was published in Kwiatkowski et al. (2014). All models showed skills in simulating some variables, but simpler models were broadly closer to observations overall. As for the optimal level of complexity, this is an unsettled question in the field of marine ecosystem modelling (e.g. Allen et al., 2010). Extremely simple models are easy to interpret but may not be able to reproduce realistic behaviour, while too much complexity will lead to uncertainty and problems in interpretation of the model (Allen et al., 2010). Given differences we observed in the studied region, our results suggest caution on relying on a single "ultimate" model for understanding regime shifts behaviour and rather favour multiple lower to intermediate complexity models, as also recommended by Fulton et al. (2003). However, one should be careful transferring these results to other regions. More complex models could outperform simple models in different ecosystems, as they have been suggested to be more portable (i.e. ability to perform well in diverse regions and physical settings) in a modelling comparative study focusing on the equatorial Pacific and Arabian Sea (Friedrichs et al., 2007). In future work, an ensemble approach to quantify the effects of model and internal variability uncertainty in regime shift detection would be beneficial. We added a section in the fourth paragraph of the discussion and conclusion section to discuss these points.

References:

Allen, J. I., Aiken, J., Anderson, T. R., Buitenhuis, E., Cornell, S., Geider, R., Haines, K., Hirata, T., Holt, J., Le Quéré, C., Hardman-Mountford, N., Ross, O. N., Sinha, B., and While, J.: Marine ecosystem models for earth systems applications: The MarQUEST experience. Journal of Marine Systems, 81, 19-33, 2010.

Friedrichs, M. A., Dusenberry, M. J., Anderson, L., Armstrong, R., Chai, F., Christian, J., Doney, S., Dunne, J., Fujii, M., Hood, R., McGillicuddy, D., Moore, M., Schartau, M., Spitz, Y., and Wiggert, J.: Assessment of skill and portability in regional marine biogeochemical models: the role of multiple plankton groups. Journal of Geophysical Research, 112, C08001, 2007.

Fulton, E. A., Smith, A. D. M., and Johnson, C. R.: Effect of complexity on marine ecosystem models. Marine Ecology Progress Series, 253, 1-16, 2003.

Kwiatkowski, L., Yool, A., Allen, J. I., Anderson, T. R., Barciela, R., Buitenhuis, E. T., Butenschön, M., Enright, C., Halloran, P. R., Le Quéré, C., de Mora, L., Racault, M.-F., Sinha, B., Totterdell, I. J., and Cox, P. M.: iMarNet: an ocean biogeochemistry model inter-comparison project within a common physical ocean modeling framework. Biogeosciences, 11, 7291-7394, 2014.

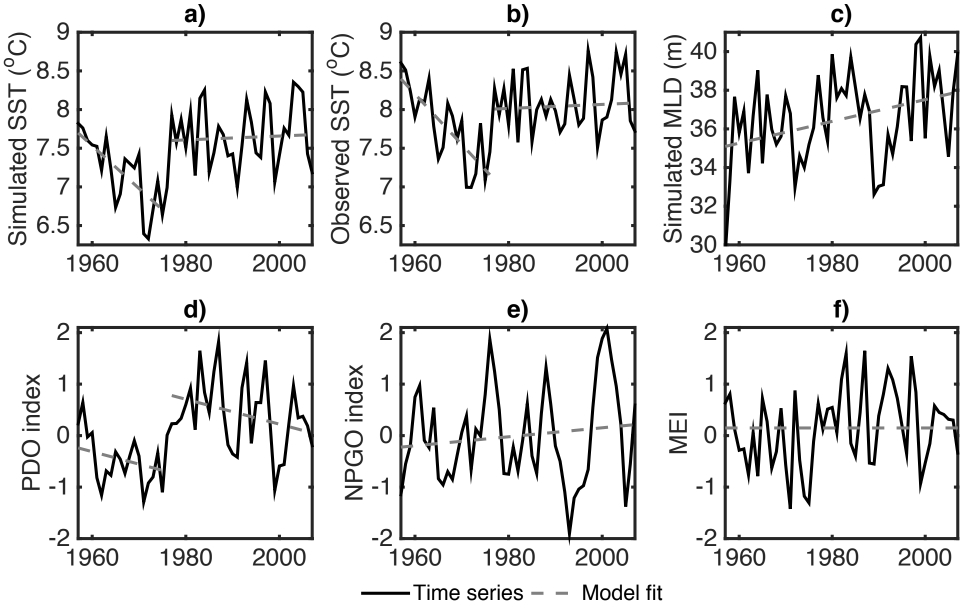

Figure 3. Time series of (a) simulated sea surface temperature (SST), (b) observed SST and (c) simulated mixed layer depth (MLD) for the Gulf of Alaska. The simulated time series of SST and MLD are the same in the five ocean models used. Time series of large-scale oscillations representing the climate in the Gulf of Alaska: (d) Pacific Decadal Oscillation (PDO) index, (e) North Pacific Gyre Oscillation (NPGO) index and (f) Multivariate El Niño Southern Oscillation index (MEI). The grey dotted lines represent the statistical model chosen (see Table A1) to fit these time series. Both the simulated SST and observed SST exhibit a significant shift in intercept and trend occurring in 1976 (p-value < 0.05, see Table A1). The MLD time series does not exhibit a significant shift and is best represented by a linear trend. Among the large-scale oscillations, only the PDO exhibits a significant shift in 1976.

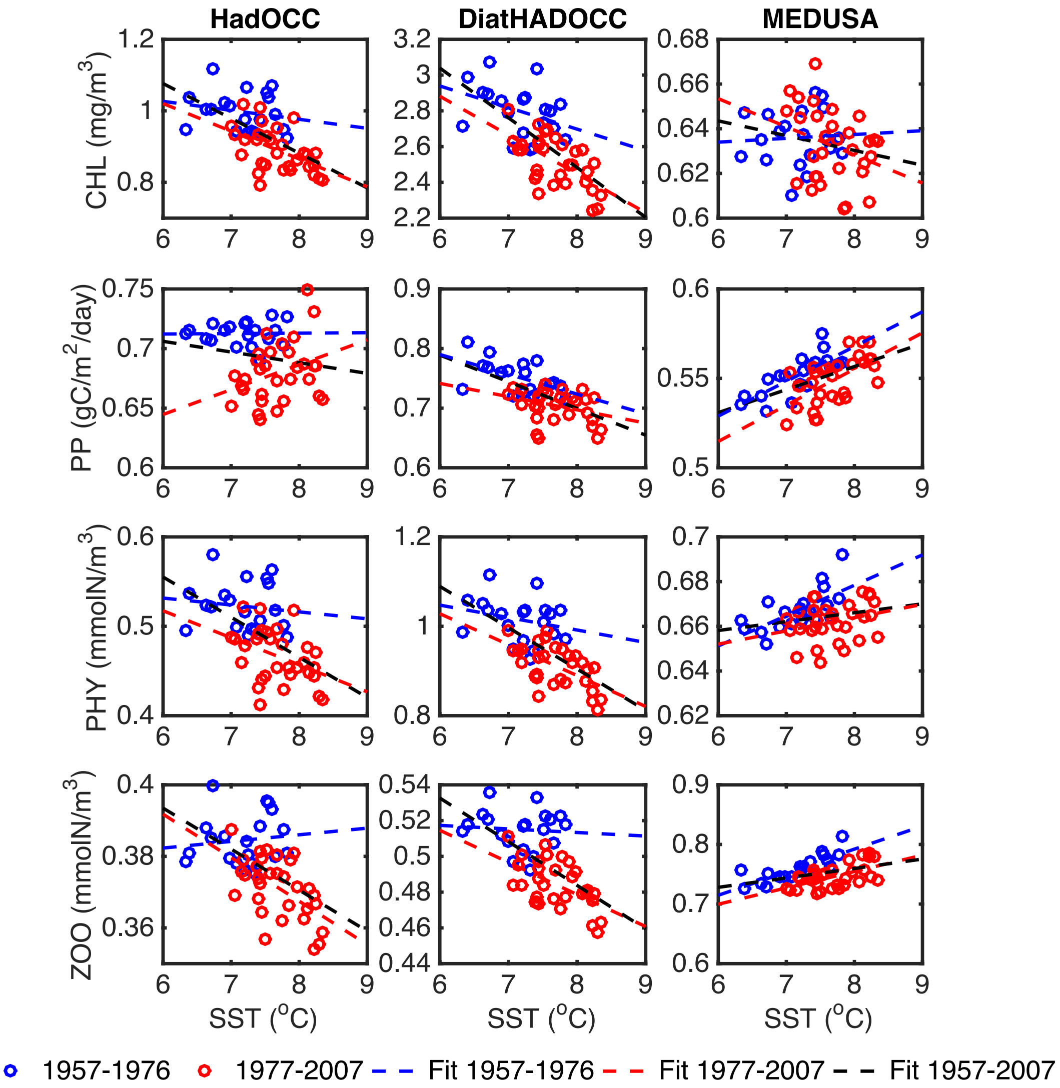

Figure 6. Relationships matrix between simulated sea surface temperature (SST) and the biological variables over the Gulf of Alaska region. Columns represent different models (HadOCC, DiatHadOCC and MEDUSA) and rows represent different biological variables (surface chlorophyll (CHL), integrated primary 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.

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. The shifts detected here are different from the red noise in the sense that they will be significant if unlikely to occur (given the strength of the red noise and trend in the background). This is already an improvement over several regime shift detection methods not making this distinction. Of course, a simultaneous estimation of shifts/red noise/trend may still confuse these parameters, but this is 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 red noise are appropriate. However, this assumption may not be appropriate at a smaller time scale (e.g. monthly). In response to this valuable comment, we now express this caveat in the discussion (last paragraph).
Our results rather show that the relationship between SST and most biological variables (except zooplankton) is similar after the shift, suggesting a linear response as opposed to an amplified nonlinear response to small changes in forcing (see Table 5). Therefore, our results do not contradict Rudnick and Davis who states that detection of regime shifts is "not evidence of nonlinear processes leading to bi-stable behavior".

(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 (red noise), which is especially tricky to visually interpret as it creates signals and patterns that will look like shifts, especially in short time series. The autocorrelation 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 these figures in the revised version (Figure 6), 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.

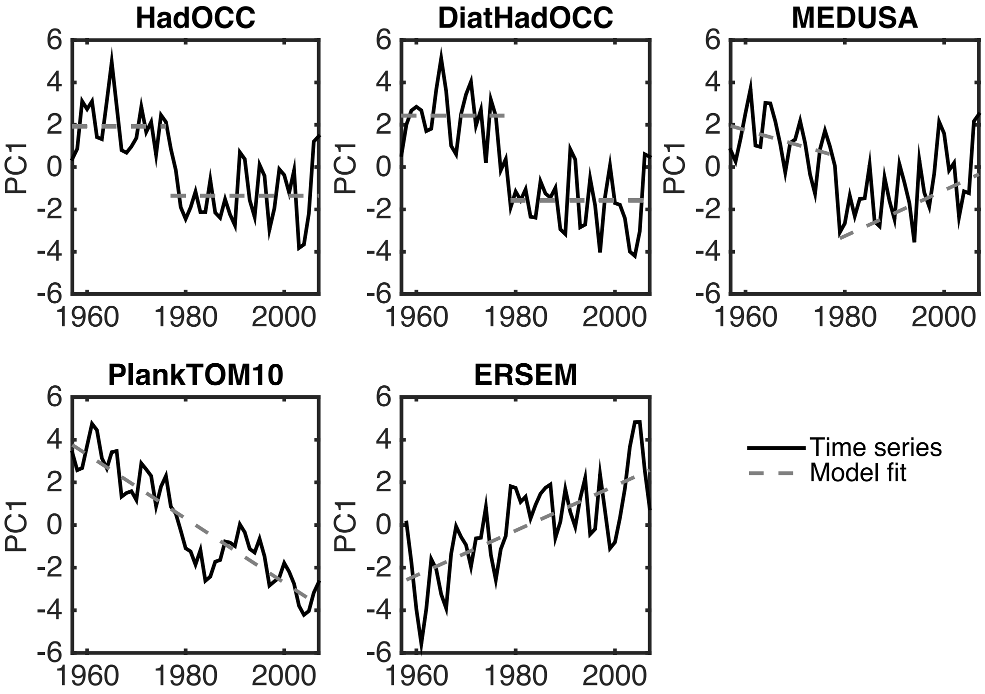

Figure 4. First principal component (PC1) of sea surface temperature, mixed layer depth, surface dissolved inorganic nitrogen, silica, iron, surface chlorophyll, integrated primary production, total surface phytoplankton and zooplankton biomass (if available) averaged over the Gulf of Alaska for each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

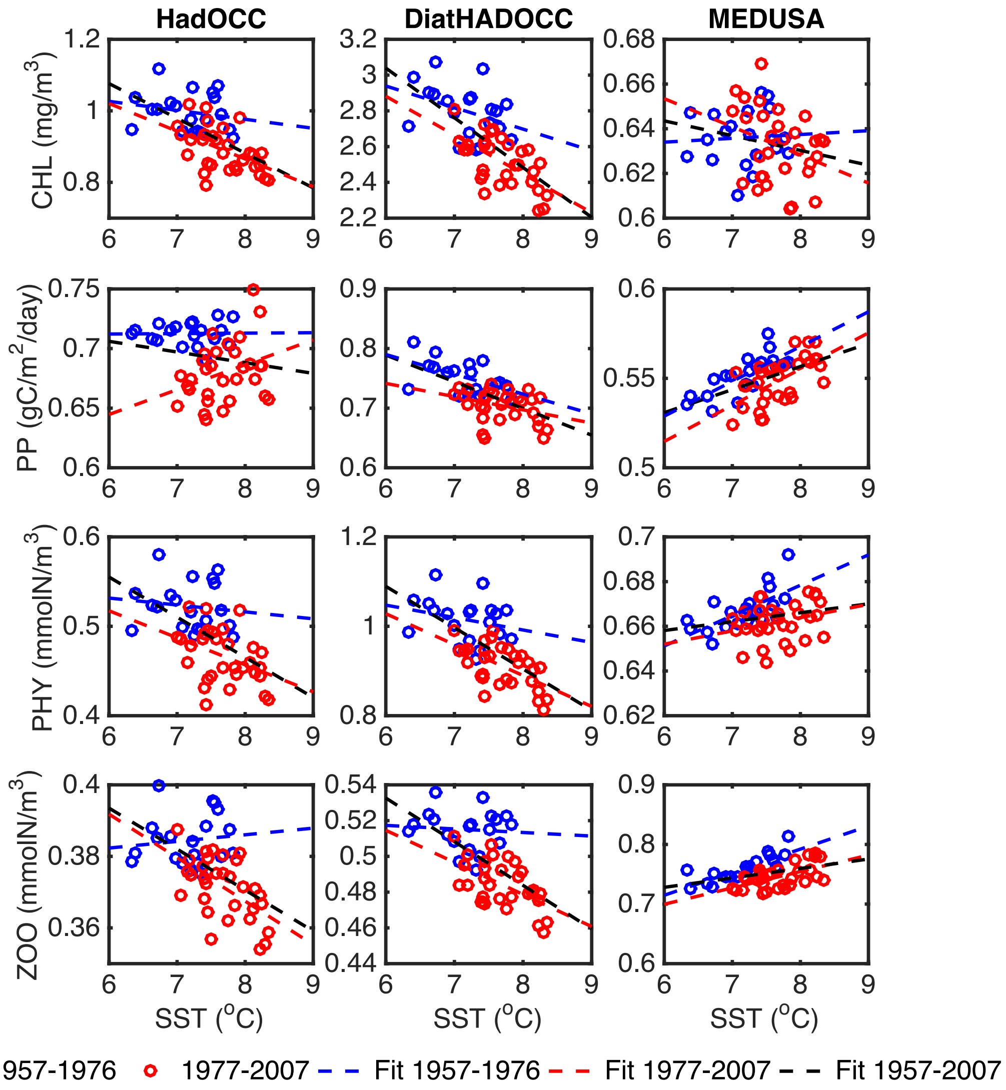

Figure 6. Relationships matrix between simulated sea surface temperature (SST) and the biological variables over the Gulf of Alaska region. Columns represent different models (HadOCC, DiatHadOCC and MEDUSA) and rows represent different biological variables (surface chlorophyll (CHL), integrated primary 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.

Response to anonymous referee 3:

This manuscript aims to apply a new method to the detection of regime shifts in time series of biological and physical variables. The Gulf of Alaska was selected as a test case for presentation of this method. I like the overall approach that the author has taken with his analysis, which presents a more rigorous method for identification of shifts in time series, including the filtering out of more red noise type events. Unfortunately I found too many grammatical errors, and odd word usage or confusing sentences which detracted from the enjoy-ability of this paper to suggest that it is ready to publish as is. I have tried to highlight an example of some of these in my specific comments but revision should be carefully checked for grammar. Additionally, I did not get the feeling that the author was overly familiar with the existing literature on regime shifts and ecosystem dynamics in the Gulf of Alaska as there was no discussion about either the Pacific Decadal Oscillation (PDO) or the ENSO which operates on a shorter time scale than the PDO but could be equally important in driving ecosystem dynamics in this region. If the manuscript was revised to address these concerns I think it could be a valuable contribution to the literature.
Specific comments: {} - odd word choice [] - suggested replacement

We highly appreciate the comments of this referee, especially for the time he/she took to make suggestions to improve the language, which make the manuscript clearer. There are native English speakers in the list of authors, who have carefully checked the revised version to correct grammar mistakes.

Furthermore, following the referee suggestion, we added a discussion on these drivers of the North Pacific climate, along with the North Pacific Gyre Oscillation (discussion and conclusions section, second paragraph). We further analyzed these indices, and stress previous results suggesting a significant shift in the PDO in the late 1970s.

<u>Page 14004</u>

L14 ..Our study demonstrates that ocean biogeochemical models are capable of simulating the late 1970s shift, {indicating} [manifested as] an abrupt increase in sea surface temperature forcing followed by an abrupt decrease in nutrients and biological productivity.

We have replaced “indicating” with “manifested as”

L20....the 1980s shift was {constrained} [driven by] by changes in mixed layer

We have replaced “constrained” with “driven by”

L22.....simulate regime shifts in the Gulf of Alaska region, {thereby} [These models can therefore be considered useful tools to enhance our] {providing a better} understanding..

We have modified the sentence as suggested by the referee.

Page 14005

L8. The regime shift can be. . .. This is a run-on/confusing sentence that needs rewording

We have split the sentence into two parts to avoid confusion.

L18. . .respond differently to [an ecosystem subject only to] natural disturbances

We modified this sentence as suggested.

L20 . It might be better to replace 'global-warming' with climate change.

We replaced "global-warming" with "climate change"

L22. . . shifts in bottom-up control in the food web via phytoplankton or zooplankton. This needs to be expanded on. Presumably you mean through temporal or spatial changes to the lower trophic level carbon production but this is not clear.

We have clarified the sentence and now specify " via changes affecting the abundance of phytoplankton or zooplankton"

L32 – This sentence sound repetitive of the one preceding it. I would suggest combining the two.

We are unsure which sentence the referee is referring to here as there is no L32 in page 14005.

Page 14006

L6 –A substantial part. . . This information doesn't seem to fit here. I would move this whole paragraph to the end of the introduction – just prior to Paragraph 5 where you describe organizational flow.

The paragraph was moved as suggested by the referee.

L6- Multivariate principle component analysis can tell us how the different components of the ecosystem are responding together. I think this is what you are doing but you should say as much to inform the reader not familiar with this type of analysis.

We apply principal component analysis to the z-scores of the physical and biological time series averaged over the Gulf of Alaska for each model to reduce the dimensions of all variables analysed into uncorrelated principal components. We also apply the change-point methodology to the first principal component (PC1) obtained for each model, which explains most of the variability, and test whether PC1 also exhibit a shift in the late 1970s. We then investigate which variables are contributing most to the late 1970s shift, by comparing their individual contributions to PC1 for each model. We added these details about principal component in the methodology section.

Page 14007

L1 {forcings} I would replace with [events]

We would rather stick with "forcings" here as it is the term also used in other studies we are referring to.

L3..and [able to] distinguish these. . .

We replaced "distinguish" with "distinguishing" instead

L7..and threw response {can be} [was] explored..

We replaced "can be " with "was" as suggested by the referee

L24.. [More} specifically[,] in the Gulf of Alaska, a modeling study..

We removed "Specifically" as suggested.

L27.Increases in spring zooplankton biomass were observed.. Over what time period?

We specify the increase was observed when comparing the periods of 1956-1962 and 1980-1989.

Page 14008

L2 – over what time period were the observed increases?

From the 1970s to the 1990s, as mentioned in the following sentence

L7 – Inconsistent use of tense. ..climate shift occurred. . .ecological response varie{s}[d]

Verb tense corrected as suggested by the referee.

L8 – Further south – Than what? Presumably the GOA but it is not clear in this paragraph.

Than the Gulf of Alaska – now specified.

L12..not a return to pre 1977.. So what were the changes ? would be good to inform the reader of this.

As suggested by another referee, discussion of the late 1980s shift has been removed from the manuscript and so was this sentence.

L22. It is unlikely that we will be able to have long term predictions of regime shifts that correspond to the timing of an actual regime shift. You would not expect the forward looking climate models that are used to driving the ocean models to capture the timing of the regime shifts. The may be able to predict regimes in a statistical sense in that they have the right number occurring over the right time frame i.e. decadal, but all models of this nature generally have limited success simulating the timescales of variability and the chaotic randomness (internal variations) found in the real world and should not be expected to align temporally with a model's internal variations (Walston et al., 2014).

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.

L29 to be {described} [assigned] only to their representation. .

We replaced described by ascribed instead, as suggested by referee 2.

<u>Page 14010</u>

L3 – {errors} due to the physics —I don't think it's correct to talk about errors- unless you are assessing which of a suite of physical models is better. I would just call them [differences]

We replaced "errors" with "differences" as suggested.

L9 ..dissolved oxygen [initial condition] fields. . .

We added initial condition.

<u>Page 14012</u>

L5 – Does the CORE forcing have a 'shift' in the forcing? Presumably it does and this is what is driving the shift in ocean temperature that you see. Is CORE a re-analysis product that assimilates temperature observations? If so we would expect it to reflect reality and the timing of the regime shift. If not, it would be quite surprising that it managed to simulate the correct timing for the shift.

Good point. Yes we would expect it to reflect reality and the timing of the regime shift if it comes from atmospheric forcing: “CORE2 provides observationally derived geographical fields of atmospheric properties (temperature, humidity, wind), as well as downwelling heat and freshwater fluxes.” These additional details were added in the simulation description section (2.2).

<u>Page 14013</u>

L17 For the models with [a detectable] shift..

We instead added “a” before shift

<u>Page 14014</u>

L5 – if a model with [a] shift.

We added “a” before shift as suggested.

L9 – {can be} [was] added. . .

We replaced “can be” with “was”

L11 – The explanation of the Monte Carlo simulation that starts on line 21 needs to be incorporated here. Otherwise it is not at all clear what you are talking about.

We believe we need to first introduce the concepts of decision rule and critical value to explain what we are estimating using Monte Carlo simulations. Thus, we kept the order as is.

<u>Page 14016</u>

Results are usually presented in past tense

We thoroughly checked that the tense we use to present the results is consistent.

<u>Page 14017</u>

L20 –Even though our analysis.. This sentence needs re-wording for clarity.

The sentence was reworded as: “Even though our analysis does not suggest a significant shift in MLD in the late 1970s, a subtle change is suggested by the cumulative sums smooth change of slope. Similarly, a slight change of slope in MLD is observed in the late 1980s.”

<u>Page 14018</u>

L26 – There is one {possible} exception .. It is an exception, not a possible one.

We removed “possible”.

L29 {forcing} [driver]

We replaced “forcing” with “driver”

Page 14019

L1 {Therefore} the forcing. . .

We removed “therefore”

L6. . .suggesting a linear [biological] response

We prefer not to add biological here as it is already mentioned in this sentence – we think it is obvious we refer to a biological response here.

Page 14019

In general there appeared to be quite a bit of repetition of the results here and less inference of what the results can tell us, why this may be significant for the Gulf of Alaska ecosystem and how this compares to others finding in this area.

We agree. We cut repetitions in the discussion section and added a comparison to the results of other modelling studies (e.g. Haigh et al., 2001; Capotondi et al., 2005; Alexander et al., 2008) - third paragraph in the revised version of the manuscript. Our results are in agreement with Haigh et al. (2001), who are suggesting a year-round deepening of the mixed layer depth in the Gulf of Alaska after 1976, which led to a decrease in nutrient, phytoplankton and zooplankton after 1976. Alexander et al. (2008) rather simulate a shoaling in the winter mixed layer depth in the late 1970s, giving rise to a early-spring increase in primary production, phytoplankton and zooplankton biomass followed by a late-spring decline in both phytoplankton and zooplankton biomass. Despite the caveat that we are analysing annual mean time series it is important to point out the contradictory direction of change in mixed layer depth. Possibly reconciling this discrepancy, Capotondi et al. (2005) suggest a deepening trend in a broad band along the coast and shoaling in the central part of the Gulf of Alaska. Thus, the comparison of the various attempts to simulate the late 1970s regime shift of the Gulf of Alaska raises the possibility that the abrupt and spatially coherent ecosystem change is actually caused by a previously unappreciated heterogeneous set of environmental changes with distinct spatial pattern and timing in the annual cycle. Further analysis would be required to investigate changes at the seasonal scale and at a finer spatial resolution and is beyond the scope of this study.

L16 – would you expect this shift given the forcing you used to drive the models?

We would not necessarily expect this shift as sea surface temperature was not relaxed to match the observations but sea surface salinity was weakly relaxed (characteristic timescale of 180 days) towards observations to minimize drift. This has been added in the simulation description in section 2.2.

Page 14021

L14 – The fact that Polovina used observations and found a shallowing mixed layer post regime in this region but the models are all predicting a deepening warrants a more in-depth discussion. Are the models all wrong? What use are they if they can simulate regime shifts but with the wrong response?

See response above for p.14019

Page 14023

L2- Predictability. . .See earlier comment about predictability of regime shifts

A mentioned earlier, we agree with the referee that the prediction of regime shifts is a different problem and removed any mention of it in the manuscript.

References:

Alexander, M., A. Capotondi, A. Miller, F. Chai, R. Brodeur and C. Deser, 2008: Decadal variability in the Northeast Pacific in a physical-ecosystem model: The role of mixed layer depth and trophic interactions. Journal of Geophysical Research - Oceans, 113, C02017, doi:10.1029/2007JC004359.

Capotondi, A., M. A. Alexander, C. Deser, and A. J. Miller (2005), Low frequency pycnocline variability in the northeast Pacific, J. Phys. Oceanogr., 35, 1403 – 1420.

Haigh, S. P., K. L. Denman, and W. W. Hsieh (2001), Simulation of the planktonic ecosystem response to pre- and post-1976 forcing in an isopycnic model of the North Pacific, Can. J. Fish. Aquat. Sci., 58, 703 – 722.

Marine regime shifts in ocean biogeochemical models: a case study in the Gulf of Alaska

C. Beaulieu[1], H. Cole[1,*], S. Henson[2], A. Yool[2], T. R. Anderson[2], L. de Mora[3], E. T. Buitenhuis[4], M. Butenschön[3], I. J. Totterdell[5], J. I. Allen[3]

[1] {Ocean and Earth Science, University of Southampton, Southampton, UK}

[2] {National Oceanography Centre, Southampton, UK}

[3] {Plymouth Marine Laboratory, Plymouth, UK}

[4] {Tyndall Centre for Climate Change research, School of Environmental Sciences, University of East Anglia, Norwich, UK}

[5] {Hadley Centre, Met Office, Exeter, UK}

[*] {Now at: MSS Marine Laboratory, Marine Scotland Science, Aberdeen, UK}

Correspondence to: C. Beaulieu (c.beaulieu@soton.ac.uk)

Abstract

Regime shifts have been reported in many marine ecosystems, and are often expressed as an abrupt change occurring in multiple physical and biological components of the system. In the Gulf of Alaska, a regime shift in the late 1970s was observed, indicated by an abrupt increase in sea surface temperature and major shifts in the catch of many fish species. A thorough understanding of the extent and mechanisms leading to such regime shifts is challenged by data paucity in time and space. We investigate the ability of a suite of ocean biogeochemistry models of varying complexity to simulate regime shifts in the Gulf of Alaska by examining the presence of abrupt changes in time series of physical variables (sea surface temperature and mixed layer depth), nutrients and biological variables (chlorophyll, primary productivity and plankton biomass) using change-point analysis. Our results show that some ocean biogeochemical models are capable of simulating the late 1970s shift, manifested as an abrupt increase in sea surface temperature followed by an abrupt decrease in nutrients and biological productivity. Models from low to intermediate complexity simulate an abrupt transition in the late 1970s (i.e. a significant shift from one year to the next) while the transition is smoother in higher complexity models. Our study demonstrates that ocean biogeochemical models can successfully simulate regime shifts in the Gulf of Alaska region. These models can therefore be considered useful tools to enhance our understanding of how changes in physical conditions are propagated from lower to upper trophic levels.

1 Introduction

Although there is no universal definition of a marine regime shift, they are typically described as an abrupt change in the ecosystem from one state to another, which is detectable in multiple physical and biological components of the system (Lees et al., 2006; Daskalov et al., 2007; deYoung et al., 2008; Andersen et al., 2009; Schwing, 2009). Generally, the magnitude of the regime shift is large and it occurs rapidly relative to the time spent in the different states (e.g. a shift from one year to the next that persists on decadal or longer time scales). The regime shift can be a linear response to an abrupt change in forcing (e.g. climate shift), a nonlinear response to a small change in forcing or driven by the internal dynamics of the system (Andersen et al., 2009; Bestelmeyer et al., 2011), but the exact mechanisms are often unknown.

Key drivers of marine regime shifts include changes in ecosystem habitat, biotic processes such as dynamics of the foodweb and abiotic processes such as changes in physical and chemical conditions (deYoung et al., 2008). These drivers can be natural or anthropogenic, or a combined influence, which can increase the vulnerability of ecosystems (e.g. an ecosystem which has less resilience due to increasing human pressure tends to respond differently to an ecosystem subject only to natural disturbances) (Folke et al., 2004). Excessive fishing is an example of an anthropogenic biotic driver where a decrease in top predators (top-down control) can cause a trophic cascade, resulting in a new bottom-up controlled state (Daskalov et al., 2007). Abiotic factors such as climate change or ocean and atmosphere oscillations may initiate bottom-up regime shifts in the food web via changes affecting the abundance of phytoplankton or zooplankton (Cury and Shannon, 2004). Typically, bottom-up driven shifts in biological components of the ecosystem generated by climate shifts manifested in changes in sea surface temperature or mixed layer depth are considered the most easily identified (deYoung et al., 2008) and are the focus of this study.

Temporal and spatial scales of regime shifts may also affect their detectability (e.g. from a small scale coral reef regime shift occurring within a single year to a North Pacific - wide ecosystem regime shift taking a few years to transition) (Drinkwater, 2006; deYoung et al., 2008). Hence, detection of a shift in a large complex marine ecosystem such as the North Pacific or North Atlantic, in which there may be lags between the expression of the shift in the abiotic and biotic components of the system, may be more difficult than detecting a regime shift in a small coral reef (deYoung et al., 2008).

Regime shifts associated with changes in physical conditions have been previously reported in the North Atlantic (Drinkwater, 2006; Beaugrand et al., 2009; Alheit et al., 2014), North Sea (Reid et al., 2001; Beaugrand, 2004; McQuatters-Gollop et al., 2007) and North Pacific (Polovina et al., 1995; Mantua et al., 1997; Hare and Mantua, 2000; Litzow and Mueter, 2014), among others. The late 1970s North Pacific regime shift has been comprehensively studied (Mantua et al., 1997; McGowan et al., 1998; Francis et al., 1998; Hare and Mantua, 2000; Yatsu et al., 2008). It was observed in a composite time series of 100 physical and biological variables, which revealed an abrupt and sustained change during 1976/77 (Hare and Mantua, 2000). At that time, there was a deepening of the Aleutian low pressure system which doubled the eastward wind stress and brought cooler winds over the central North Pacific, causing a drop in sea surface temperature (SST) and a deepening of the mixed layer depth (MLD). This resulted in moister and warmer air settling over the California Current region and the Gulf of Alaska, which caused an increase in SST in these two regions (Mantua et al., 1997). This mechanism has been described as the Pacific Decadal Oscillation (PDO), which switched from a negative to a positive state in 1976/77 (Mantua et al., 1997). The late 1970s shift is thereby implicitly related to El Niño Southern Oscillation (ENSO) variability, whose shorter timescale fluctuations combined with random atmospheric forcing enforce decadal variability in the PDO (Newman et al., 2003). Alternatively, other large-scale climate patterns such as the North Pacific Gyre Oscillation (NPGO) may impact on marine ecosystem dynamics. Concurrent with the switch in the PDO state, an increase in zooplankton biomass was observed in the Gulf of Alaska between the periods of 1956-1962 and 1980-1989 (Brodeur and Ware, 1992). In upper trophic levels, abrupt increases in groundfish recruitment and salmon catches were observed, while some forage fish populations collapsed with consequences for piscivorous sea birds and marine mammal populations (Anderson and Piatt, 1999). Overall the yield of fish stocks in the Gulf of Alaska increased from the 1970s to the 1990s (McGowan et al., 1998).

Although a climate shift occurred over the entire North Pacific, the ecological response varied between regions depending on their respective dominant processes (Schwing, 2009). For example, in the California Current region the ecological changes associated with the 1977 climate shift were different from those that occurred in the Gulf of Alaska with lower salmon catches after 1977 (Mantua et al., 1997). Investigation of the magnitude and extent of the regime shift and the proposed mechanism is challenged by the paucity of data covering adequate time and space scales in the Gulf of Alaska. Most support for the observed biological changes comes from fisheries stock assessments, which are not designed to study how climate shifts are affecting marine ecosystems (McGowan et al., 1998). A few modelling studies have attempted to simulate the chain of events for the late 1970s shift, but the direction of changes in the simulations of the physical and biological parameters are sometimes opposite, and also vary according to the space/time scale of the study (e.g. Polovina et al., 1995; Haigh et al., 2001; Capotondi et al., 2005; Alexander et al., 2008). By using the late 1970s regime shifts in the Gulf of Alaska as a case study we aim to assess the ability of five global ocean biogeochemical models to simulate this shift. The models were part of the UK Integrated Global Biogeochemical Modelling Network (iMarNet) intercomparison, which aimed to evaluate the models' ability to simulate global-scale bulk biogeochemical properties using the same ocean general circulation model and atmospheric forcing (Kwiatkowski et al., 2014). These physically identical hindcast simulations allow any model differences to be ascribed only to their representation of biogeochemical processes, thereby providing insight into the mechanisms leading to marine regime shifts.

A substantial part of the literature on regime shifts uses principal component analysis to compress a large number of time series representing the state of the ecosystem to a smaller number of uncorrelated ones, which indicates to what extent the different components of the system are responding coherently. For example, Hare and Mantua (2000) reduced a total of 100 time series of physical and biological variables representing the state of the North Pacific to two leading modes of variability. The presence of regime shifts in the reduced set of time series may render the presence of shifts more evident to visual inspection, but this is often done without further significance testing (Andersen et al., 2009). In order to objectively identify the timing of a shift and distinguish it from a random fluctuation, change-point techniques can be used, especially methods designed to detect multiple shifts in the mean of a time series (e.g. Andersen et al., 2009). For example, the shift detection methodology proposed by Rodionov (2004) consists of applying a t-test successively to compare the means of two segments of a time series, considering all possible timings for a shift, and repeats this until all shifts have been detected. This method has been applied widely in the marine regime shift literature (e.g. Daskalov et al., 2007; DeYoung et al., 2008; Overland et al., 2008; Yatsu et al., 2008; Möllmann et al., 2009; Overland et al., 2010). However, it is not designed to distinguish a shift from a trend, which may lead to the detection of a series of spurious shifts in the presence of a background long-term trend (e.g. Spencer et al., 2011). Furthermore, it may lead to the detection of spurious shifts in the presence of red noise, which creates patterns that may be interpreted as shifts, but which are purely random (e.g. Wunsch, 1999). Red noise is often present in biological time series such as chlorophyll (e.g. Beaulieu et al., 2013) or plankton abundance (e.g. Di Lorenzo and Ohman, 2013), and manifests through a slow integrated response to random weather forcings (Di Lorenzo and Ohman, 2013). Therefore, we opt for a methodology capable of separating a long-term trend from an abrupt change signal (e.g. which occurs from one year to the next) and distinguishing these signals from red noise (Beaulieu et al., 2012). In order to provide further insights as to whether the shifts detected are a linear response to a shift in the forcing itself (e.g. climate shift) from shifts generated through a nonlinear response of some change in the forcing, also called thresholds or "tipping points" (Scheffer et al., 2009), the relationship between the forcing and the response was explored using regression models (Bestelmeyer et al., 2011).

Our analysis is organised as follows. First, we investigate whether shifts are present in the Gulf of Alaska as predicted in a multiple model intercomparison hindcast experiment, iMarNet (Kwiatkowski et al., 2014; imarnet.org). Specifically, we analyse model physical and biological variables for regime shifts and verify whether these shifts are internally coherent. Then, we investigate the contribution of the different physical and biological variables to the observed late 1970s and late 1980s shifts in the Gulf of Alaska and the type of forcing-response relationship that led to abrupt changes.

2 Methodology

2.1 Ocean biogeochemical models

This study uses ocean biogeochemistry model (OBGC) outputs from the iMarNet intercomparison project. The primary aim of iMarNet was to investigate the model complexity required to adequately represent marine ecosystems (Kwiatkowski et al., 2014). The participating models were HadOCC (Palmer and Totterdell, 2001), Diat-HadOCC (Halloran et al., 2010), MEDUSA-2 (Yool et al., 2011; 2013), PlankTOM10 (Le Quéré et al., 2005) and ERSEM (Baretta et al., 1995; Blackford et al., 2004). These models cover a large span of model complexity from 7 state variables (including 2 plankton functional types; PFTs) in HadOCC through to 57 state variables (including 8 PFTs) in ERSEM. The hindcast simulations (covering the period 1957 to 2007) from each of the models were used in this study.

The key focus of the iMarNet intercomparison was to evaluate the models' ability to simulate global-scale bulk properties, such as carbon and nutrient cycles, as a representation of marine biotic activity (Kwiatkowski et al., 2014). The different OBGC models were implemented within a common physical framework to eliminate confounding differences due to the physics that would otherwise occur if different physical models were involved. This framework used the Nucleus for European Modelling of the Ocean (NEMO) physical ocean general circulation model (Madec, 2008) coupled to the Los Alamos sea-ice model (CICE; Hunke and Lipscomb, 2008), with surface atmospheric forcing drawn from the common ocean-ice reference experiment (CORE2; Large & Yeager, 2009). The model grid was configured at approximately 1°-degree horizontal resolution, with 75 vertical levels increasing in thickness from 1m at the surface to 200m at 6000m depth.

The models were initialised from an identical physical state in 1890 using the same 3D biogeochemical tracer fields (although not all of these tracers were used in every model). Macronutrients (nitrate, phosphorus, silicic acid) and dissolved oxygen initial condition fields were drawn from the World Ocean Atlas 2009 (Garcia et al., 2010a, 2010b), while fields of dissolved inorganic carbon and alkalinity were drawn from the Global Ocean Data Analysis Project (GLODAP) database (Key et al., 2004). Each model used its own source for iron fields as currently there is no comprehensive global dataset available. The remaining fields such as plankton and particulate and dissolved organic matter were initialized with arbitrary small initial conditions. Below is a brief description of the structure of each OBGC model, which is also summarised in Table 1. Additional details can be found in Kwiatkowski et al. (2014).

- The Hadley Centre Ocean Carbon Cycle (HadOCC) model is a simple NPZD (Nutrient, Phytoplankton, Zooplankton, Detritus) model consisting of one phytoplankton group and one zooplankton group. There is one nutrient pool, nitrogen, to which the cycling of carbon and alkalinity is coupled. Further details can be found in Palmer and Totterdell (2001).
- Diat-HadOCC is a descendant of HadOCC with the primary difference being the presence of 2 phytoplankton groups: diatoms and mixed phytoplankton. Further differences include the addition of the nutrients silica and iron and the effect of nutrient limitation on growth is multiplicative, where light limitation is multiplied by successive nutrient limitation terms. Further details can be found in Halloran et al. *(*2010).
- Model of Ecosystem Dynamics, nutrient Utilization, Sequestration and Acidification (MEDUSA) is an intermediate complexity model comprising two phytoplankton and two

zooplankton groups. The ecosystem is split into small (nanophytoplankton and microzooplankton) and large (diatom and mesozooplankton) components, and non-living detrital material is similarly split to reflect its sources. Nutrient pools included in this model are nitrogen, silica and iron and the effect of nutrient limitation on growth is also multiplicative. Cycles of carbon, alkalinity and dissolved oxygen are also included. Further details can be found in Yool et al. (2011) and Yool et al. (2013).

- PlankTOM10 is a relatively complex model and has 10 PFTs (diatoms, coccolithophores, *Phaeocystis*, nitrogen fixers, picophytoplankton, mixed phytoplankton, protozoa, mesozooplankton, macrozooplankton and bacteria). The nutrient cycles included in PlankTOM10 are carbon, nitrogen, oxygen, phosphorous, silica and a simplified iron cycle. Phytoplankton growth is regulated by the minimum of nutrient limitation terms. All zooplankton groups eat smaller PFTs, with preference based on size. Further details can be found in Le Quéré et al. (2005) and Buitenhuis et al. (2013).
- The European Regional Seas Ecosystem Model (ERSEM) was originally used for shelf seas and consists of both pelagic and benthic ecosystems. Four phytoplankton groups (picophytoplankton/flagellates, flagellates, large phytoplankton and diatoms), three zooplankton groups (heterotrophic flagellates, microzooplankton and mesozooplankton) and heterotrophic bacteria are represented. Each zooplankton group grazes on a preferred phytoplankton group or groups based on size. The nutrient pools consist of carbon, nitrogen, phosphorous, silica and dissolved oxygen allowing for dynamic stoichiometric internal quotas. The effect of nutrient limitation on growth is a combination of multiplicative and maximum limitation factors. More details can be found in Blackford (1997), Blackford et al. (2004) and Butenschön et al. (2015).

2.2 Simulation

For each biogeochemical model, conventional simulations from the same physical initial state were performed identically from year 1890 through to 2007. For the first 60 years of these simulations (1890-1949 inclusive), CORE2 (Common Ocean-ice Reference Experiments, version 2; Large and Yeager, 2009) seasonal climatology (i.e. without interrannual variability) was used, the so-called "normal year forcing". Subsequently (1950-2007 inclusive), interannually-varying CORE2 forcing was used to complete the simulations. CORE2 provides observationally-derived geographical fields of downwelling irradiance (short- and long-wave), precipitation (rain and snow), air temperature, humidity, and meridional and zonal winds. These are used in conjunction with bulk formulae to calculate net heat, freshwater and momentum exchange between the atmosphere and the ocean. In addition, sea surface salinity was weakly relaxed (characteristic timescale of 180 days) towards observations to minimise drift. Note that the simulations were "online", in that physics and biogeochemistry were both formally simulated simultaneously. Feedbacks between the model biology and ocean physics (e.g. by the absorption of downwelling solar radiation) were disabled so that all of the biogeochemical models experienced consistent simulated physics. Additional details on the simulations can be found in Kwiatkowski et al. (2014).

For each model, where available, time series of sea surface temperature (SST), mixed layer depth (MLD, defined as a density difference from the surface of 0.1 kg m^{-3}), surface dissolved inorganic nitrogen (DIN), silica (SI), iron (FE), surface chlorophyll (CHL), integrated primary production (PP), total surface phytoplankton (PHY) and zooplankton (ZOO) biomass were extracted from 1957-2007 (same period as the observational dataset used, see section below) for the Gulf of Alaska region. 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).

2.3 Observational dataset

To compare shifts found in model time series to observed ones, SST data were extracted from the Extended Reconstructed Sea Surface Temperature (ERSST) dataset (version 3b) downloaded from https://www.ncdc.noaa.gov/ersst/. This analysis uses the International Comprehensive Ocean-Atmosphere Data Set SST data and combines ship and buoy data (Smith and Reynolds, 2003; Smith et al., 2008). The data were available as monthly means with a spatial resolution of 2° x 2° from 1957 to 2007. The ERSST dataset was averaged spatially for each year over the Gulf of Alaska to form a time series of annual mean SST. Comparison with observed time series for other variables (i.e. MLD, DIN, SI, FE, CHL, PP, PHY, ZOO) is not possible due to lack of data over suitable space and time scales. Time series of large-scale oscillations representing the climate over the North Pacific were obtained. The PDO index (Mantua et al., 1997) was downloaded from http://www.atmos.washington.edu/~mantua/abst.PDO.html. The Multivariate ENSO Index (MEI; Wolter and Timlin, 1998) was downloaded from http://www.esrl.noaa.gov/psd/enso/mei/ and the NPGO index (DiLorenzo et al., 2008) was downloaded from http://www.o3d.org/npgo/npgo.php. Annual time series of PDO, ENSO and NPGO indices were produced by averaging monthly time series.

2.4 Statistical analyses

For the regime shift detection, we use the change-point detection method presented in Beaulieu et al. (2012), which distinguishes shifts in a time series from long-term trends and red noise. It consists of fitting a suite of regression models to a time series with (I) constant mean, (II) shift in the mean, (III) trend, (IV) shift in the intercept of the trend and (V) shift in both the intercept and trend, and discriminates between them. Figure 1 illustrates the five regression models tested in this study and their equations are presented in Table 2. This methodology is based on the Schwarz Information Criterion (SIC), which is a measure of goodness of fit based on the maximum likelihood function of a given model penalised by the number of parameters estimated to ensure balance between good fit and parsimony. We use the SIC to 1) identify the timing of the shift under a model formulation containing a shift and 2) determine which regression model (among the five fitted) provides the best fit. The SIC formulations for the five models are presented in Table 2. For the models with a shift (II, IV, V), the SIC is calculated for each possible timing of a shift – the timing with the lowest SIC corresponds to the year that the shift is most likely to have occurred. The search for the most likely timing for a shift excludes the first and last five data points in the time series to avoid spurious detection (Beaulieu et al., 2012). For example, the most likely timing for a shift for model II would be:

$$SIC_{II}(p) = \min\left\{SIC_{II}(k), \quad k = 5,\ldots,n-5\right\} \qquad (1)$$

The most likely timing for a shift under models IV and V can be found similarly, and are denoted $SIC_{IV}(p)$ and $SIC_{V}(p)$, respectively.

Once the SIC of the five models are computed, the smallest one is selected as the most appropriate to represent the time series (Table 2). If the SIC of a model without a shift (constant mean (I) or trend (III)) is lower than the SIC of the models with a shift (shift in the mean (II), shift in the intercept (IV) or shift in the intercept and trend (V)), no abrupt change is detected in that time series. On the other hand, if a model with a shift has the smallest SIC, this indicates that there could be a shift in that time series.

There is no significance level involved with the decision rule presented above and shifts tend to be too easily detected (Beaulieu et al., 2012). Therefore, a critical value can be added to the decision rule to assess the significance of the shift based on the difference in SIC between the shift model and the null model and is determined using Monte Carlo simulations. For example, if model II is selected with the smallest SIC, the null model to compare with is model I. The shift detected in model II will be significant if

$$SIC_{II}(p) - SIC_{I} < c_{\alpha} \qquad (2)$$

where c_{α} is the critical value at the α critical level and is determined by Monte Carlo simulation. Similarly, when models IV or V have the smallest SIC, the shift will be significant if

$$SIC_{IV}(p) - SIC_{III} < c_{\alpha} \qquad (3)$$

or

$$SIC_{V}(p) - SIC_{III} < c_{\alpha} \qquad (4)$$

We generate 1000 synthetic time series randomly drawn from a Normal distribution with the same length (i.e. number of years), variance and first-order autocorrelation (if present) as the data. The presence of autocorrelation usually indicates the presence of external factors not accounted for in the model and the AR(1) should act as a parameter which roughly comprises these factors. The SIC differences between the model with a shift (e.g. model II) and the corresponding null model (e.g. model I) are calculated. This produces a null distribution for c_{α} against which the observed SIC difference is compared to estimate the p-value. The p-value here is the probability of observing a SIC difference at least as extreme as that observed under the null hypothesis of no shift in the time series. We use a 5% critical level, i.e. we reject the null hypothesis of no shift if the p-value is smaller than 0.05. This analysis is based on the assumption that the residuals of the selected model are normally distributed with a constant variance, which is verified using a Lilliefors test and Fisher test (5% critical level) respectively. Violation of these assumptions could indicate the presence of additional shifts in the time series.

This method is flexible and allows for the detection of shifts that are more complex than simply a shift in the mean. Furthermore, it distinguishes potential shifts from red noise, which is important given the background climate change trend and long memory of the climate system (characterized as high first-order autocorrelation). However, this method can detect at most one shift in the time series, while there could possibly be multiple shifts over a multidecadal time period. Therefore, the shift identified will be the largest to occur in a time series, which for the Gulf of Alaska is expected to be the 1977 regime shift.

To unveil shifts in SST in and around the Gulf of Alaska, we first apply this methodology to observed annual SST time series over the North Pacific (from 40-70°N and 180-120°W). Second, we apply this methodology to time series of physical and biological variables simulated from each of the five ocean biogeochemical models, and to observed SST, averaged over the Gulf of Alaska as described in sections 2.1 and 2.2 respectively. As a visual aid, we also calculate cumulative sums of the z-scores of each time series. Cumulative sums are useful for monitoring time series as they exhibit a change of slope when a shift in the time series occurs (e.g. Page, 1954).

We apply principal component analysis to the z-scores of the physical and biological time series averaged over the Gulf of Alaska for each model to reduce the dimensions of all variables analysed here into uncorrelated principal components. We also apply the change-point methodology to the first principal component (PC1) obtained for each model, which explains most of the variability, and test whether PC1 also exhibit a shift in the late 1970s. We then investigate which variables are contributing most to the late 1970s shift, by comparing their individual contributions to PC1 for each model.

We further investigate the physical forcing – biological response relationship in models that simulate a significant shift in the late 1970s in PC1. We investigate the presence of changes in physical-biological relationships before and after the shift by comparing the regression slopes, following the approach proposed by Bestelmeyer et al. (2011). Similar slopes before and after the shift could indicate a linear response to the physical forcing, while a change in the slopes might rather suggest a change in the relationship and thus, a nonlinear response. More specifically, we fit simple linear regression models, such as:

$$\begin{aligned} y_t &= a_1 + b_1 x_t + e_t \quad t = 1,\ldots,p \\ y_t &= a_2 + b_2 x_t + e_t \quad t = p+1,\ldots,n \end{aligned} \qquad (5)$$

where y_t represents the biological response (either CHL, PP, PHY or ZOO), x_t is the physical forcing (either SST or MLD), a_1 and b_1 are the intercept and regression slope before the shift at time p, a_2 and b_2 are the intercept and regression slope after the shift and e_t are the white noise errors. To verify whether the relationships are similar before and after the shift, we test whether the slopes are equal ($b_1 = b_2$) using the Student test statistic (with *n-4* degrees of freedom) described by Paternoster et al. (1998):

$$t = \frac{b_1 - b_2}{s_{b_1 - b_2}} \qquad (6)$$

$$s_{b_1 - b_2} = \sqrt{s_{b_1}^2 + s_{b_2}^2} \qquad (7)$$

where b_1 and b_2 are estimated using least squares with s_{b_1} and s_{b_2} being the respective standard errors.

3 Results

Figure 2 presents the results of the change-point analysis on gridded SST observations for the North Pacific. This reveals a predominant shift in 1977 over the Gulf of Alaska region, which also extends as a coastal band towards the California Current region and the Bering Sea. A late 1980s shift is detected in a smaller area in the middle of the gyre.

In the observed SST time series averaged over the Gulf of Alaska, a statistically significant shift is detected and manifests as a rapid increase in the mean of ~1°C after a decreasing trend (Fig. 3a). In the model physical time series (which are identical in all 5 OBGC models), SST exhibits the same signal as the observations: a shift in the intercept and gradient occurring in 1976, while the MLD is best represented by a linear trend. However, the model MLD time series shows strong decadal variability with large changes occurring in the mid-1970s and at the beginning of the 1990s (Fig. 3b-c). Results of change-point analysis on large-scale oscillations characterizing the climate over the region are also presented in Fig. 3 (d-f) and show a significant shift in the PDO in the late 1970s while the NPGO and MEI annual time series do not indicate a shift.

The change-point analysis was performed on PC1 for each model (Figure 4, Table 3), which explains most of the variance for each model (except MEDUSA, 36% of variance explained) (Table 4). HadOCC exhibits a shift in 1977 in PC1 (Table 3), for which all variables except MLD have large relative contributions (>10% relative contribution, Table 4). The first principal component in Diat-HadOCC exhibits a shift in 1976 and explains 63% of the total variance. The variable offering the smallest relative contribution is again the MLD (Table 4).

In MEDUSA, a shift is also detected in the late 1970s in the first component, which explains only 36% of the variance. The SST, CHL and nutrients are the most important variables with relative contributions larger than 10% (Table 4). The MLD has a relative contribution of 0.94% to PC1 (Table 4). The relative contributions of the nutrients in the HadOCC, Diat-HadOCC and MEDUSA late 1970s shift detected in the first principal component suggests the controlling factor is nutrient limitation (i.e. bottom up control) in these models. In ERSEM and PlankTOM10, there are no shifts detected in the first principal component.

The results of the change-point analysis on all observational and model individual time series are presented in Appendix A (Table A1). The fit of the most appropriate statistical models for the biological variables for each OBGC model are also presented in Appendix A (Figs. A1-A5). Statistically significant shifts are found more often in the simpler OBGC models (HadOCC, Diat-HadOCC and MEDUSA) than the complex ones (Table A1), which is consistent with the results obtained on the first principal component for each model. Of the statistically significant shifts identified in these models, the majority occurred in the late 1970s. In HadOCC, the late 1970s shift corresponds to a decrease in DIN, CHL, PHY and ZOO, while a large increase in PP is detected in 1991. Nevertheless, PP is decreasing over the period 1957-1990 (Fig. A1). In Diat-HadOCC, all parameters exhibit a shift in the late 1970s, although it is not significant in PHY and ZOO. The significant shifts in the late 1970s manifest as a decrease in SI, FE, CHL and PP. In MEDUSA, shifts in DIN and FE (although not significant) are identified in the late 1970s. ERSEM exhibits a significant shift in CHL in the late 1970s, while PlankTOM10 does not have any significant shifts for that period.

As a visual support for the change-point analysis, cumulative sums of the z-scores of each time series within each model are presented in Fig. 5. A shift in a time series is revealed by a change of slope of the cumulative sums. The change of slope in SST is sharp, as one would expect given the significant shift detected. Even though our analysis does not suggest a significant shift in MLD in the late 1970s, a subtle change is suggested by smooth change of slope in the cumulative sum. Similarly, a slight change of slope in MLD is observed in the late 1980s. These changes are clearly propagated to the other parameters in HadOCC, DiatHadOCC and MEDUSA with a sharp change of slope, but smoother change in ERSEM and PlankTOM10.

We further investigate the forcing-response relationship between SST and the biological variables (CHL, PP, PHY, ZOO) in HadOCC, DiatHadOCC and MEDUSA (Fig. 6) before and after 1977, as a significant shift is present in PC1 in these models. The slopes of the linear relationships between SST and the biological variables are mostly similar before and after 1977 (Table 5). This is consistent with a linear, rather than nonlinear, response to changes in SST forcing. There is one exception for ZOO for which the difference in slopes is significant with a stronger relationship after 1977 in HadOCC and DiatHadOCC (Table 5, Fig. 6), which could suggest an amplified nonlinear response.

4 Discussion and Conclusions

Using the Gulf of Alaska as a case study, our results demonstrate the usefulness of OBGC models to infer the chain of events responsible for regime shifts, especially in regions where observations are scarce. Although there are many definitions of regime shifts in the literature, they can be generally described as an abrupt change (e.g. from one year to the next) that occurs across both physical and biological parts of the ecosystem. Therefore, to determine if a regime shift has occurred in the five OBGC models tested here the shift has to be traceable from physical parameters through to biological parameters. With the change-point detection method used here, we found statistically significant shifts in the late 1970s in the Gulf of Alaska simulated in five OBGC models. A shift in model SST occurred in 1976 and matched a shift in observed SST. This abrupt change in SST was accompanied by a smooth deepening of the mixed layer depth followed by an overall decrease in nutrients and productivity. The three OBGC models simulating an abrupt change in 1977 in PC1 (i.e. HadOCC, DiatHadOCC and MEDUSA) are consistent in the direction of change (Fig. 4). The decrease in nutrients after 1977 seems to be the dominant driver in the reduction in productivity and outweighs any potential advantage to phytoplankton from increased light availability that results from the shallower mixed layer. The dominance of declining nutrients in explaining the variability in the principal components of HadOCC, Diat-HadOCC and MEDUSA, supports this hypothesis.

Previous studies have linked the late 1970s shift in the North Pacific with the PDO, which switched from a negative to a positive state in 1976/77 (Mantua et al., 1997). The PDO fluctuations have been suggested to exhibit a red noise response to atmospheric noise and ENSO events (Newman et al., 2003), thereby raising the possibility of a link between ENSO and the North Pacific shift in the late 1970s. Nevertheless, the PDO (and implicitly ENSO) alone is not enough to characterize the North Pacific climate (Bond et al., 2003). Alternatively, the North Pacific Gyre Oscillation (NPGO) has been suggested as a global-scale mode of variability that plays an important role in decadal changes in marine ecosystems (DiLorenzo et al., 2008). For example, in the California Current, the PDO correlates with SST while NPGO is more closely related to variability in salinity, nutrient and primary production (DiLorenzo et al., 2008). Thus, if both the PDO and NPGO fluctuations drive changes in the North Pacific climate and ecosystem functioning, the question arises whether either or both of these indices exhibit a shift at a similar time. Underscoring some of the conclusions of the prior work discussed above, the shift in 1976/77 manifests in the PDO index, but notably we find no significant shifts in the multivariate ENSO index or the NPGO index. Clearly, by detecting a shift in the late 1970s in PDO only we cannot conclusively tie the PDO and untie the NPGO and ENSO to the shift in climate and ecosystem dynamics of the Gulf of Alaska. However, these corresponding changes are an important piece of information to future work aimed at determining causal mechanisms, mode of teleconnection and coupled physical/biogeochemical dynamics that link global climate patterns to ocean fertility of the Gulf of Alaska.

In conclusion, the 1977 regime shift in the Gulf of Alaska was observed in sea surface temperature and in the abundance of a range of commercial fish species (McGowan et al., 1998). Here, we infer the behaviour of the nutrients and lower trophic levels using OBGC models, and the relationship of these changes to physical variables that are plausible drivers. Our novel approach based on change-point detection offers a helpful framework to evaluate previous modelling studies that have attempted to reproduce the extent of changes from physics to biology for the late 1970s shift in the Gulf of Alaska (e.g. Polovina et al., 1995; Haigh et al., 2001; Capotondi et al., 2005; Alexander et al., 2008). For example, Haigh et al. (2001) used the Miami isopycnic coordinate ocean model combined with an ecosystem model of 4 compartments (Denman and Peña, 1999) to show that a year-round deepening of the mixed layer depth after 1976 led to a slight decrease in nutrients and phytoplankton as well as zooplankton biomass. These findings are broadly consistent with the model simulations analysed here. Other studies instead suggest that the MLD shoaled after 1977 resulting in increased plankton production in the region. This is the case in the Polovina et al. (1995) study, which suggested that shoaling in the spring/winter MLD led to increased productivity in a plankton population dynamics model. More recently, Alexander et al. (2008) used the National Center for Atmospheric Research Climate System Model Ocean Model (NCOM) combined with a biological model that contains 10 compartments (Chai et al., 2002) to simulate the chain of events in the region. In that study, an increase in SST simulated in the late 1970s is accompanied by a shoaling in the winter mixed layer depth, giving rise to an early spring increase in primary production, phytoplankton and zooplankton biomass followed by a late spring decline in both phytoplankton and zooplankton biomass. Despite the caveat that we are analysing annual mean time series it is important to point out the contradictory direction of change in mixed layer depth. Possibly reconciling this discrepancy, Capotondi et al. (2005) suggest, based on NCOM model simulations, a deepening trend in MLD in a broad band along the coast and shoaling in the central part of the Gulf of Alaska. Thus, the comparison of the various attempts to simulate the late 1970s regime shift of the Gulf of Alaska raises the possibility that the observed abrupt and spatially coherent ecosystem change was actually caused by a previously unappreciated heterogeneous set of environmental changes with distinct spatial pattern and timing in the annual cycle. If so, the inherent assumption underpinning our own and previous work to understand the Gulf of Alaska ecosystem shift as a single mechanistic causal sequence may be overly simplistic. Consequently, future analysis aimed at spatial and temporal heterogeneity of abrupt regional ecosystem change has the potential to greatly improve our understanding of the underlying dynamics and the vulnerability of marine ecosystems to abrupt future changes.

A second major outcome of this study involves the role of model complexity in determining a system’s propensity for abrupt ecosystem change. All the OBGC models used in this study have the same underlying physical model, and were run with the same initial conditions and forcing fields. Their performance in terms of a fit to observations has been assessed globally in a previous study by Kwiatkowski et al. (2014), showing that all models have skills in simulating some variables, but simpler models were broadly closer to observations overall. In the Gulf of Alaska, the five models systematically differ in nutrient and biological responses as a function of model ecosystem complexity. Simple to intermediate complexity models such as HadOCC, Diat-HadOCC, and MEDUSA simulate a shift in the late 1970s, which manifests as an abrupt change in SST and many nutrients and biological parameters. As the model complexity increases to PlankTOM10 and ERSEM, these changes are mostly in the same direction but become less abrupt. The simpler models have fewer plankton groups responding to environmental changes (both HadOCC and Diat-HadOCC have one zooplankton group, and Diat-HadOCC has two phytoplankton groups), which might explain a more direct response than a model with a larger number of plankton groups interacting with each other. More complex models could potentially unveil shifts in the community structure (i.e. increase of a certain type of plankton and decrease of another one), as regime shifts can affect different species in opposite ways (Benson and Trites, 2002). Feedbacks and interactions between groups in the models are in need of thorough exploration to determine how they affect the simulation of observed regime shifts. Such differences between model results raise the question as to what degree of model complexity is needed to appropriately simulate the complexity of regime shifts in the real world. Extremely simple models are easy to interpret but may not be able to reproduce realistic behaviour, while too much complexity will lead to uncertainty and problems in interpretation of the model (Allen et al., 2010). Given the observed differences between models, our results suggest caution on relying on a single “ultimate” model for understanding regime shifts behaviour and rather favour multiple lower to intermediate complexity models, as recommended by Fulton et al. (2003). However, our results should not be generalised too easily, as we focused uniquely on the Gulf of Alaska region here. More complex models could outperform simple models in different ecosystems. For example, higher complexity models have been suggested to be more portable (i.e. ability to perform well in diverse regions and physical settings) in a comparative study focusing on the equatorial Pacific and Arabian Sea (Friedrichs et al., 2007). Future work should involve a regime shift analysis in several ecosystems using models with traceable complexity. Furthermore, an ensemble approach to quantify the effects of model and internal variability uncertainty in regime shift detection would be beneficial.

Our analysis suggests that the Gulf of Alaska regime shift is consistent with a linear response to physical forcings on lower trophic levels, showing a bottom-up response due to changes in the physical environment controlled via nutrient limitation, with a potential amplified response from ZOO (only in HadOCC and Diat-HadOCC). This result is in agreement with the linear tracking window hypothesis (Hsieh and Ohman, 2006), which suggests that some populations can respond linearly to abrupt changes in physical forcing, as opposed to an amplified nonlinear response to small changes in forcing (e.g. Scheffer et al., 2009). However, it must be noted that our analysis is lacking top-down controls from upper trophic levels beyond zooplankton, and thus only partly resolves possible explanations for the observed regime shifts in the Gulf of Alaska. Many drivers (and their synergistic effects) may combine to fully explain regime shifts (Lindegren et al., 2012; Litzow et al., 2014). Models including upper trophic levels able to simulate regime shifts would also be beneficial to better understand the mechanisms leading to the shift and estimate critical thresholds.

Finally, beyond model complexity and the spatial and temporal resolution at which the output is analysed, the state-of-the-art in statistical techniques for regime shift detection is an active area of research. Here we employ an approach to detect shifts and distinguish them from a long-term trend and background red noise, i.e. evaluate if the shift is unusually large given the fluctuations that would be expected in the presence of autocorrelation and/or a trend (Beaulieu et al., 2012), which is an improvement over previous methodologies. A main current limitation of this methodology is the ability to detect at most one shift and in one time series at a time (univariate), but work to extend the methodology to detect multiple shifts in a multivariate setting is under way. Furthermore, we distinguish against a background of red noise, which is assumed constant through the time series, but the presence of changes in the red noise through time could affect the results. For example, a recent study suggests a "reddening" of the PDO and North Pacific SST as an explanation for occurrences of abrupt changes in the North Pacific ecosystem (Boulton and Lenton, 2015). However, this is unlikely to affect our results given the time scale (annual means) and length of the time series (51 years) used in this study. Further, we suggest here that analysis of the forcing-response relationship helps to distinguish between a regime shift with a linear response to a shift in forcing, and a nonlinear response after crossing a forcing threshold, as originally proposed by Bestelmeyer et al. (2011). Here we used a test that is based on a quantitative comparison of the forcing-response relationship before and after the shift. This approach can be used to detect other marine or terrestrial regime shifts and distinguish between a linear and a nonlinear response to external forcing. For management purposes, distinguishing between these two types of forcing-response relationship producing regime shifts is critical, as they will lead to different management and policy incentives (Kelly et al., 2015). For example, a routine monitoring of threshold-based systems leads to better management outcomes than "threshold-blind" management, i.e. when ignoring the possibility of a threshold and assuming a linear forcing-response relationship (Kelly et al., 2015).

Appendix A

This appendix presents the results of the change-point analysis for all parameters simulated from the five models. The physical parameters (SST and MLD) are omitted here as they are presented in Fig. 3. The chosen model for each variable and each OBGC model is presented in Table A1.

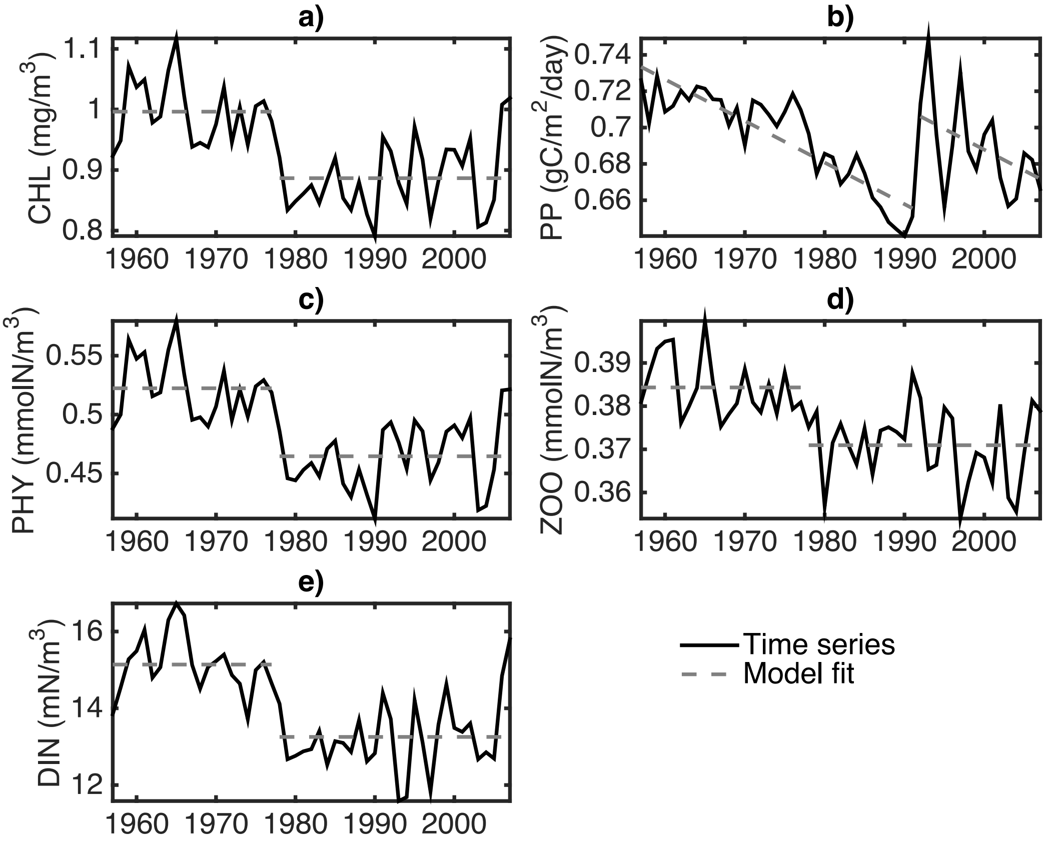

Figure A1. Time series of a) surface chlorophyll, b) integrated primary production, c) total surface phytoplankton, d) zooplankton biomass and e) surface dissolved inorganic nitrogen simulated with the **HadOCC** model and averaged over the Gulf of Alaska region. The dotted lines represent the statistical model selected.

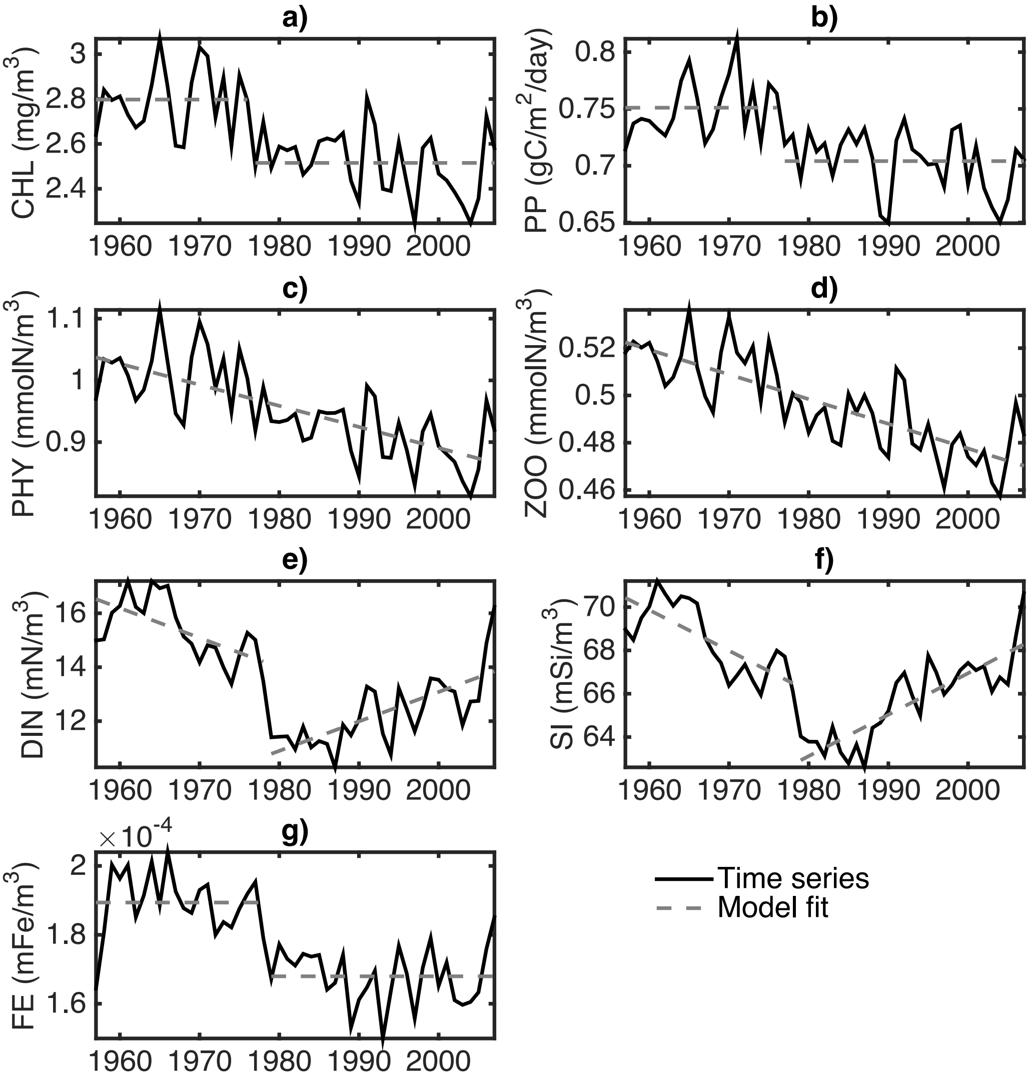

Figure A2. Time series of a) surface chlorophyll, b) integrated primary production, c) total surface phytoplankton, d) zooplankton biomass and e) surface dissolved inorganic nitrogen, f) silica and g) iron simulated with the **DiatHadOCC** model and averaged over the Gulf of Alaska region. The dotted lines represent the statistical model selected.

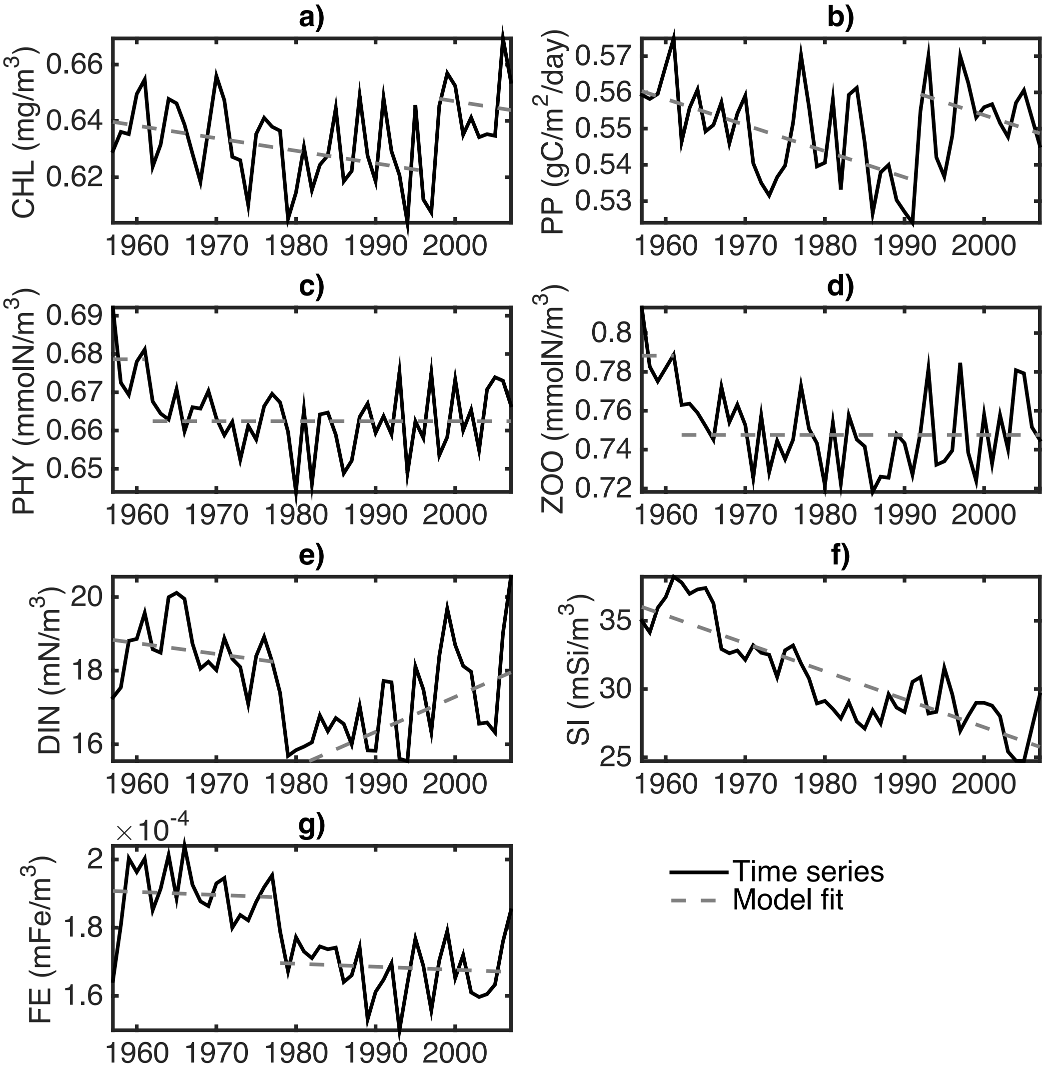

Figure A3. Time series of a) surface chlorophyll, b) integrated primary production, c) total surface phytoplankton, d) zooplankton biomass and e) surface dissolved inorganic nitrogen, f) silica and g) iron simulated with the **MEDUSA** model and averaged over the Gulf of Alaska region. The dotted lines represent the statistical model selected.

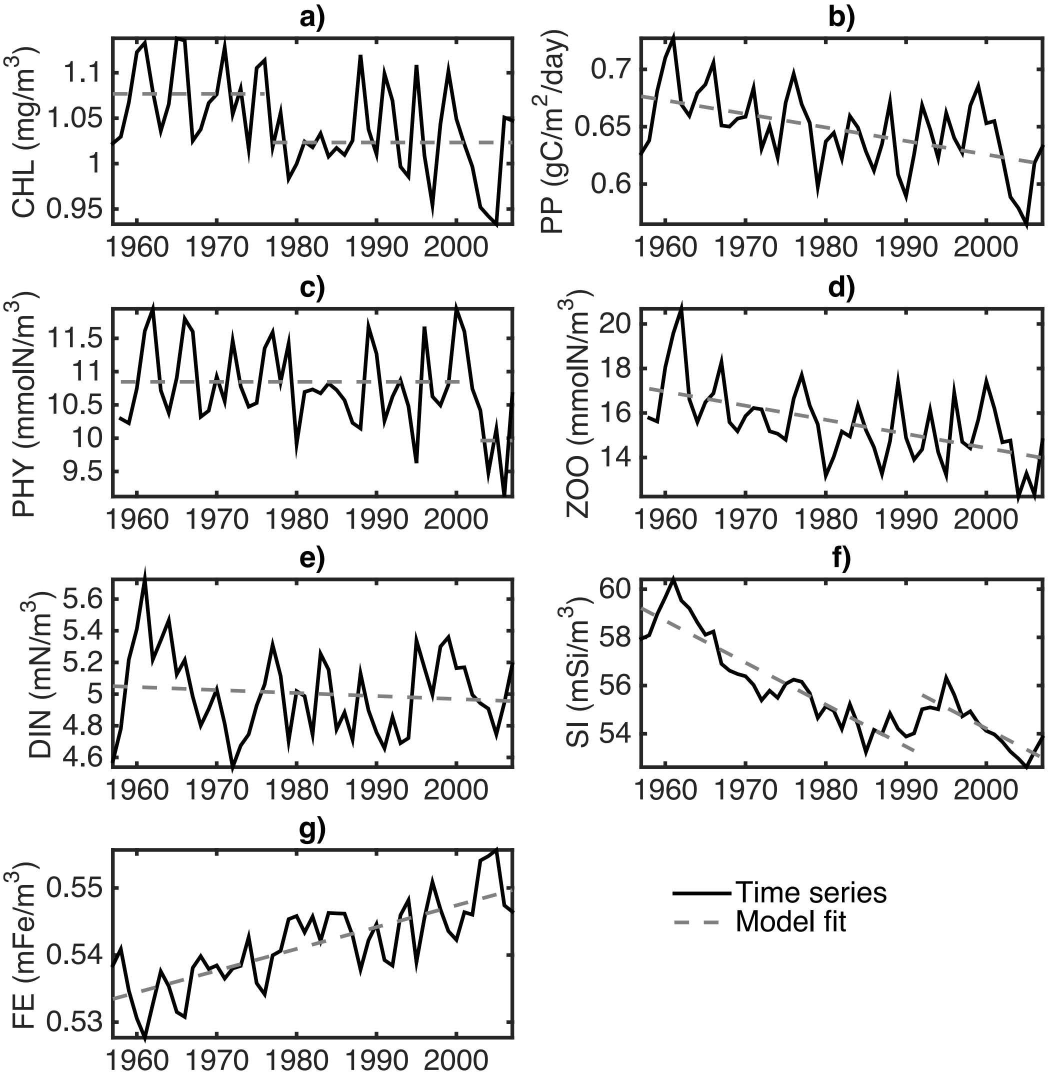

Figure A4. Time series of a) surface chlorophyll, b) integrated primary production, c) total surface phytoplankton, d) zooplankton biomass and e) surface dissolved inorganic nitrogen, f) silica and g) iron simulated with the **PlankTOM10** model and averaged over the Gulf of Alaska region. The dotted lines represent the statistical model selected.

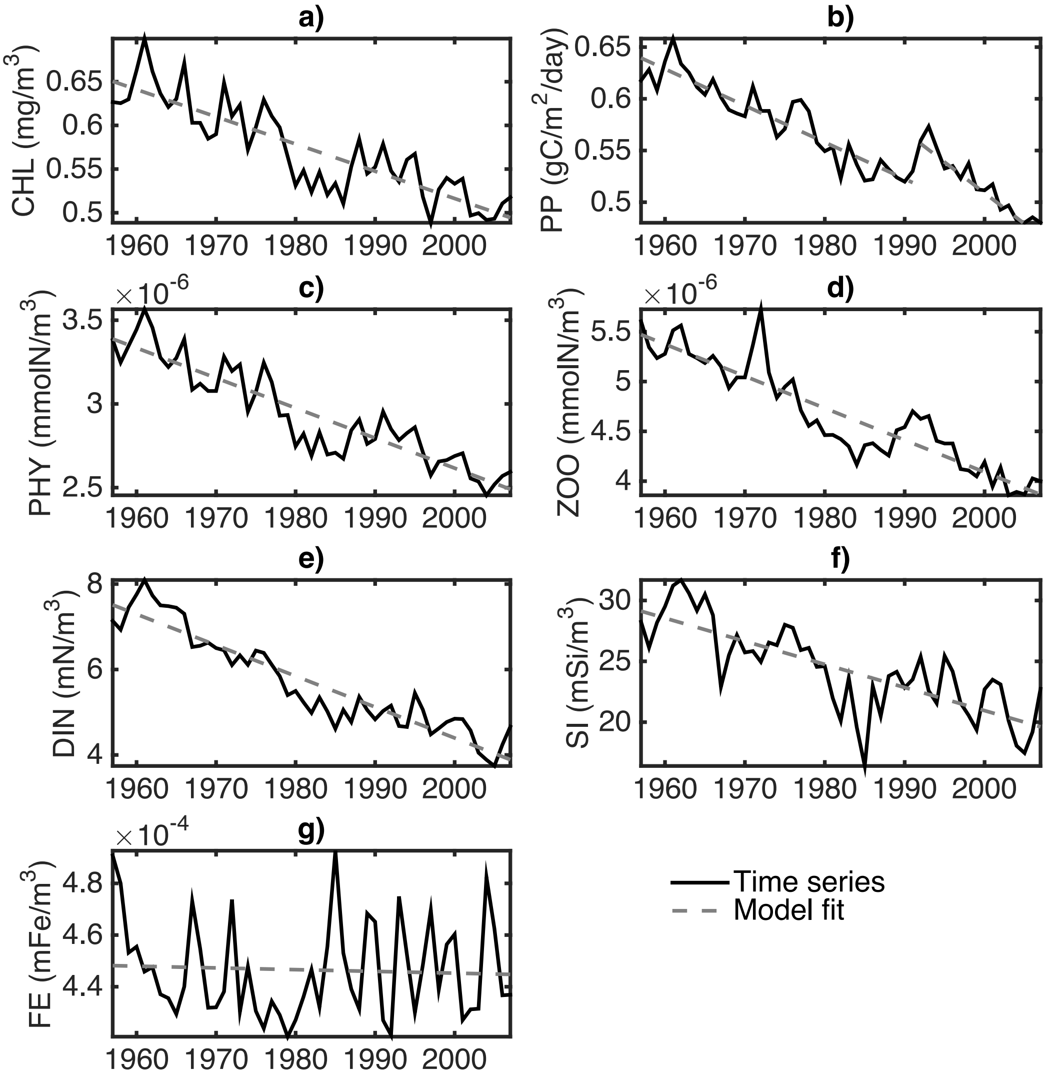

Figure A5. Time series of a) surface chlorophyll, b) integrated primary production, c) total surface phytoplankton, d) zooplankton biomass and e) surface dissolved inorganic nitrogen, f) silica and g) iron simulated with the **ERSEM** model and averaged over the Gulf of Alaska region. The dotted lines represent the statistical model selected.

Table A1. Results from change-point detection analysis for all observational and modelled time series. Years in bold have a significant shift (p-value < 0.05).

	Parameter	Shift year	Shift type	SIC	SIC (Null model)	p-value
Observations	SST	**1976**	trend and intercept	52.79	70.63	<0.01
	PDO	**1976**	intercept	104.94	120.11	0.02
	ENSO	1976	mean	116.04	120.42	0.11
	NPGO	1998	trend and intercept	137.79	147.12	0.28 [c]
All models	SST	**1976**	trend and intercept	58.39	74.15	<0.01
	MLD	1987	intercept	230.22	234.25	0.25
HadOCC	CHL	**1977**	mean	-138.40	-108.32	<0.01 [c]
	PP	**1991**	intercept	-264.06	-235.87	<0.01 [b, c]
	PHY	**1977**	mean	-211.46	-177.59	<0.01 [c]
	ZOO	**1977**	mean	-339.68	-315.70	<0.01
	DIN	**1977**	mean	139.52	175.85	<0.01 [c]
DiatHadOCC	CHL	**1976**	mean	-44.93	-13.82	<0.01
	PP	**1976**	mean	-216.45	-190.71	<0.01 [c]
	PHY	1976	intercept	-157.13	-155.59	0.53
	ZOO	1976	intercept	-298.90	-297.33	0.59
	DIN	**1978**	trend and intercept	151.10	202.7	<0.01 [c]
	SI	**1978**	trend and intercept	167.04	230.11	<0.01 [c]
	FE	**1978**	mean	-1035.5	-990.86	<0.01 [c]
MEDUSA	CHL	**1997**	intercept	-287.1	-274.71	0.01
	PP	**1991**	intercept	-308.90	-293.98	0.02 [c]
	PHY	**1961**	mean	-342.52	-328.88	<0.01
	ZOO	**1961**	mean	-260.89	-243.23	<0.01
	DIN	**1978**	trend and intercept	157.02	180.64	<0.01 [c]
	SI	1966	trend and intercept	201.11	217.83	0.09 [c]
	FE	1977	intercept	-946.48	-938.51	0.10 [c]
PlankTOM10	CHL	1978	intercept	-221.06	-214.48	0.24 [c]
	PP	**1991**	trend and intercept	-277.74	-258.29	<0.01 [b, c]
	PHY	1986	intercept	-1481.6	-1472.22	0.18 [c]

	ZOO	1988	intercept	-1427.8	-1414.98	0.16 [a, b, c]
	DIN	1978	trend and intercept	48.07	62.65	0.07 [c]
	SI	1987	intercept	233.68	240.68	0.29 [c]
	FE	1983	intercept	-960.84	-954.91	0.12 [a, c]
ERSEM	CHL	**1976**	mean	-162.70	-151.07	0.01 [c]
	PP	1961	trend and intercept	-211.38	-207.73	0.49 [b, c]
	PHY	**2002**	mean	95.40	101.6	0.04
	ZOO	1961	trend and intercept	175.98	185.44	0.07 [c]
	DIN	1964	trend and intercept	6.58	16.48	0.10 [c]
	SI	**1991**	intercept	122.52	153.74	0.01 [c]
	FE	1986	intercept	-414.51	-412.18	0.57 [c]

[a] residuals not normally distributed (Lilliefors test, 5% critical level)

[b] residual variance not constant (Fisher test, 5% critical level)

[c] residuals not independent (Durbin-Watson test, 5% critical level): the Monte Carlo simulations to estimate the p-value incorporates the first-order autocorrelation of the resduals.

Acknowledgements

This work was funded by the UK Natural Environmental Research Council Integrated Marine Biogeochemical Modelling Network to Support UK Earth System Research (iMarNet) project (NE/K0011345/1). CB also acknowledges financial support from a Marie Curie Career Integration Grant (project 631466 – TROPHYZ). We acknowledge use of the MONSooN system, a collaborative facility supplied under the Joint Weather and Climate Research Programme, a strategic partnership between the Met Office and the Natural Environment Research Council. The authors would like to thank Clare Enright for providing model runs. The authors would also like to thank Matthew Spencer and Adrian Martin for useful comments on an earlier version of the manuscript, as well as three anonymous reviewers for constructive comments that improved the manuscript substantially.

References

Alexander, M., Capotondi, A., Miller, A., Chai, F., Brodeur, R., and Deser, C.: Decadal variability in the northeast Pacific in a physical-ecosystem model: Role of mixed layer depth and trophic interactions. Journal of Geophysical Research, 113, C02017, 2008.

Alheit, J. P., Licandro, S., Coombs, A., Garcia, A., Giráldez, A., Garcia Santamaría, M. T., Slotte, A., and Tsikliras, A. C.: Atlantic Multidecadal Oscillation (AMO) modulates dynamics of small pelagic fishes and ecosystem regime shifts in the eastern North and Central Atlantic. Journal of Marine Systems, 131, 21-35, 2014.

Allen, J. I., Aiken, J., Anderson, T. R., Buitenhuis, E., Cornell, S., Geider, R., Haines, K., Hirata, T., Holt, J., Le Quéré, C., Hardman-Mountford, N., Ross, O. N., Sinha, B., and While, J.: Marine ecosystem models for earth systems applications: The MarQUEST experience. Journal of Marine Systems, 81, 19-33, 2010.

Andersen, T., Carstensen, J., Hernandez-Garcia, E., and Duarte, D. M.: Ecological thresholds and regime shifts: approaches to identification. Trends in Ecology & Evolution, 24, 49-57, 2009.

Anderson, P. J. and Piatt, J. F.: Community reorganization of the Gulf of Alaska following ocean climate regime shift. Marine Ecology Progress Series, 189, 117-123, 1999.

Baretta, J.W., Ebenhoh, W., and Ruardij, P.: The European Regional Seas Ecosystem Model, a complex marine ecosystem model. Netherlands Journal of Sea Research, 33, 233–246, 1995.

Beaugrand, G.: The North Sea regime shift: evidence, causes, mechanisms and consequences. Progress in Oceanography, 60, 245-262, 2004.

Beaugrand, G, Luczak C, and Edwards, M.: Rapid biogeographical plankton shifts in the North Atlantic Ocean. Global Change Biology, 15, 1790-1803, 2009.

Beaulieu, C., Chen, J., and Sarmiento, J. L.: Change-point analysis as a tool to detect abrupt climate variations. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 370, 1228-1249, 2012.

Beaulieu, C., Henson, S.A., Sarmiento, J. L., Dunne, J. P., Doney, S. C., Rykaczewski, R. R., and Bopp, L.: Factors challenging our ability to detect long-term trends in ocean chlorophyll. Biogeosciences, 10, 2711–2724, 2013.

Bellwood, D. R., Hugues, T. P., Folke, C., and Nyström, M.: Confronting the coral reef crisis. Nature, 429, 827-833, 2004.

Benson, A. J. and Trites, A. W.: Ecological effects of regime shifts in the Bering Sea and eastern North Pacific Ocean. Fish and Fisheries, 3, 95-113, 2002.

Bestelmeyer, B. T., Ellison, A. M., Fraser, W. R., Gorman, K. B., Holbrook, S. J., Laney, C. M., Ohman, M. D., Peters, D. P. C., Pillsbury, F. C., Rassweiler, A., Schmitt, R. J., and Sharma, S.: Analysis of abrupt transitions in ecological systems. Ecosphere, 2, 129, 2011.

Blackford, J. C.: An analysis of benthic biological dynamics in a North Sea ecosystem model. Journal of Sea Research, 38, 213-230, 1997.

Blackford, J. C., Allen, J. I., and Gilbert, F. J.: Ecosystem dynamics at six contrasting sites: a generic modelling study. Journal of Marine Systems, 52, 191-215, 2004.

Bond, N. A., Overland, J. E., Spillane, M., and Stabeno, P.: Recent shifts in the state of the North Pacific. Geophysical Resaearch Letters, 30, 2183, 2003.

Boulton, C. A., and Lenton, T. M.: Slowing down of North Pacific climate variability and its implications for abrupt ecosystem change. PNAS, 112, 11496-11501, 2015.

Brodeur, R. D. and Ware, D. M.: Long-term variability in zooplankton biomass in the subarctic Pacific Ocean. Fisheries Oceanography, 1, 32-38, 1992.

Buitenhuis, E., Hashioka, T., and Le Quéré, C.: Combined constraints on global ocean primary production using observations and models. Global Biogeochemical Cycles, 27, 847-858, 2013.

Butenschön, M., Clark, J., Aldridge, J.N., Allen, J.I., Artioli, Y., Blackford, J., Bruggeman, J., Cazenave, P., Ciavatta, S., Kay, S., Lessin, G., van Leeuwen, S., van der Molen, J., de Mora, L., Polimene, L., Sailley, S., Stephens, N., and Torres, R.: ERSEM 15.06: a generic model for marine biogeochemistry and the ecosystem dynamics of the lower trophic levels. Geoscientific Model Development Discussions, 8, 7063–7187, 2015.

Capotondi, A., Alexander, M. A., Deser, C., Miller, A. J.: Low frequency pycnocline variability in the northeast Pacific. Journal of Physical Oceanography, 35, 1403-1420, 2005.

Chai, F., Dugdale, R. C., Peng, T.-H., Wilkerson, F. P., and Barber, R. T.: One-dimensional ecosystem model of the equatorial Pacific upwelling system: part I. Model development and silicon and nitrogen cycle. Deep Sea Research Part II, 49, 2713– 2745, 2002.

Cury, P. and Shannon, L.: Regime shifts in upwelling ecosystems: observed changes and possible mechanisms in the northern and southern Benguela. Progress in Oceanography, 60, 223-243, 2004.

Daskalov, G. M., Grishin, A. N., Rodionov, S., and Mihneva, V.: Trophic cascades triggered by overfishing reveal possible mechanisms of ecosystem regime shifts. PNAS, 104, 10518-10523, 2007.

Denman, K. L. and Peña, M.A.: A coupled 1-D biological/physical model of the northeast subarctic Pacific Ocean with iron limitation. Deep Sea Research Part II, 46, 2877–2908, 1999.

deYoung, B., Barange, M., Beaugrand, G., Harris, R., Perry, R. I., Scheffer, M., and Werner, F.: Regime shifts in marine ecosystems: detection, prediction and management. Trends in Ecology & Evolution, 23, 402-409, 2008.

Di Lorenzo, E., Schneider, N., Cobb, K.M., Franks, P. J. S., Chhak, K., Miller, A. J., McWilliams, J. C., Bograd, S. J., Arango, H., Curchitser, E., Powell, T. M., and Rivière, P.: North Pacific Gyre Oscillation links ocean climate and ecosystem change. Geophysical Research Letters, 35, L08607, 2008.

Di Lorenzo, E. and Ohman, M. D.: A double-integration hypothesis to explain ocean ecosystem response to climate forcing. PNAS, 110, 2496-2499, 2013.

Drinkwater, K. F.: The regime shift of the 1920s and 1930s in the North Atlantic. Progress in Oceanography, 68, 134-151, 2006.

Folke, C., Carpenter, S., Walker, B., Scheffer, M., Elmqvist, T., Gunderson, L., and Holling, C. S.: Regime shifts, resilience, and biodiversity in ecosystem management. Annual Review of Ecology, Evolution and Systematics, 35, 557–581, 2004.

Francis, R. C., Hare, S. R., Hollowed, A. B., and Wooster, W. S.: Effects of interdecadal climate variability on the oceanic ecosystems of the NE Pacific. Fisheries Oceanography, 7, 1-21, 1998.

Friedrichs, M. A., Dusenberry, M. J., Anderson, L., Armstrong, R., Chai, F., Christian, J., Doney, S., Dunne, J., Fujii, M., Hood, R., McGillicuddy, D., Moore, M., Schartau, M., Spitz, Y., and Wiggert, J.: Assessment of skill and portability in regional marine biogeochemical models: the role of multiple plankton groups. Journal of Geophysical Research, 112, C08001, 2007.

Fulton, E. A., Smith, A. D. M., and Johnson, C. R.: Effect of complexity on marine ecosystem models. Marine Ecology Progress Series, 253, 1-16, 2003.

Garcia, H. E., Locarnini, R. A., Boyer, T. P., Antonov, J. I., Baranova, O. K., Zweng, M. M., and Johnson, D. R.: World Ocean Atlas 2009, Volume 3: Dissolved Oxygen, Apparent Oxygen Utilization, and Oxygen Saturation (ed Levitus S), 344 pp., NOAA Atlas NESDIS 70, US Government Printing Office, Washington DC, 2009.

Garcia, H. E., Locarnini, R. A., Boyer, T. P., Antonov, J. I., Baranova, O. K., Zweng, M. M., and Johnson, D. R.: World Ocean Atlas 2009, Volume 4: Nutrients (Phosphate, Nitrate, Silicate) (ed Levitus S), 398 pp., NOAA Atlas NESDIS 71, US Government Printing Office, Washington DC, 2009.

Haigh, S. P., Denman, K. L., and Hsieh, W. W.: Simulation of the planktonic ecosystem response to pre- and post- 1976 forcing in an isopycnic model of the North Pacific. Canadian journal of Fisheries and Aquatic Sciences, 58, 703-722, 2001.

Halloran, P., Bell, T., and Totterdell, I.: Can we trust empirical marine DMS parameterisations within projections of future climate? Biogeosciences, 7, 1645-1656, 2010.

Hare, S. and Mantua, N. J.: Empirical evidence for North Pacific regime shifts in 1977 and 1989. Progress in Oceanography, 47, 103-145, 2000.

Hasselmann, K.: Stochastic climate models. 1. Theory. Tellus, 28, 473–48, 1976.

Hsieh, C.–H. and Ohman, M. D.: Biological responses to environmental forcing: the linear tracking window hypothesis. Ecology, 87, 1932-1938, 2006.

Hunke, E. C. and Lipscomb, W. H.: CICE: The Los Alamos Sea Ice Model, Documentation and Software User's Manual, Version 4.0, Los Alamos National Laboratory Tech. Rep. LA-CC-06, 2008.

Kelly, R. P., Erickson, A. L., Mease, L.A., Battista, W., Kittinger, J. N., and Fujitta, R.: Embracing thresholds for better environmental management. Philosophical Transactions of the Royal Society B: Biological Sciences, 370, 20130276, 2015.

Key, R. M., Kozyr, A., Sabine, C. L., Lee, K., Wanningkhof, R., Bullister, J. L., Feely, R. A., Millero, F. J., Mordy, C., and Peng, T.-H.: A global ocean carbon climatology: results from Global Data Analysis Project (GLODAP). Global Biogeochemical Cycles, 18, GB4031, 2004.

Kwiatkowski, L., Yool, A., Allen, J. I., Anderson, T. R., Barciela, R., Buitenhuis, E. T., Butenschön, M., Enright, C., Halloran, P. R., Le Quéré, C., de Mora, L., Racault, M.-F., Sinha, B., Totterdell, I. J., and Cox, P. M.: iMarNet: an ocean biogeochemistry model intercomparison project within a common physical ocean modeling framework. Biogeosciences, 11, 7291-7394, 2014.

Large, W. and Yeager, S.: The global climatology of an interannually varying air–sea flux data set. Climate Dynamics, 33, 341–364, 2009.

Lees, K., Pitois, S., Scott, C., Frid, C., and Mackinson, S.: Characterizing regime shifts in the marine environment. Fish and Fisheries, 7, 104-127, 2006.

Le Quéré, C., Harrison, S. P., Prentice, I. C., Buitenhuis, E. T., Aumont, O., Bopp, L., Claustre, H., Da Cunha, L. C., Geider, R., Giraud, X., Klaas, C., Kohfeld, K. E., Legendre, L., Manizza, M., Platt, T., Rivkin, R. B., Sathyendranath, S., Uitz, J., Watson, A. J., and Wolf-Gladrow, D.: Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models. Global Change Biology, 11, 2016-2040, 2005.

Lindegren, M., Blenckner, T., and Stenseth, N. C.: Nutrient reduction and climate change cause a potential shift from pelagic to benthic pathways in a eutrophic marine ecosystem. Global Change Biology, 18, 3491-3503, 2012.

Litzow, M. A. and Mueter, F. J.: Assessing the ecological importance of climate regime shifts: An approach from the North Pacific Ocean. Progress in Oceanography, 120, 110-119, 2014.

Litzow, M. A., Mueter, F. J., and Hobday, A. J.: Reassessing regime shifts in the North Pacific: incremental climate change and commercial fishing are necessary for explaining decadal-scale biological variability. Global Change Biology, 20, 38-50, 2014.

Madec, G.: NEMO Reference Manual, Ocean Dynamic Component: NEMO–OPA, Note du Pôle de modélisation, Technical Report 27, Institut Pierre Simon Laplace, France, ISSN No. 1288–1619, 2008.

Mantua, N. J., Hare, S. R., Zhang, Y., Wallace, J. M., and Francis, R. C.: A Pacific interdecadal climate oscillation with impacts on salmon production. Bulletin of the American Meteorological Society, 78, 1069-1079, 1997.

McGowan, J. A., Cayan, D. R., and Dorman, L. M.: Climate-ocean variability and ecosystem response in the Northeast Pacific. Science, 281, 210-217, 1998.

McQuatters-Gollop, A., Raitsos, D. E., Edwards, M., Pradhan, Y., Mee, L. D., Lavender, S. J., and Attrill, M. J.: A long-term chlorophyll data set reveals regime shift in North Sea phytoplankton biomass unconnected to nutrient trends. Limnology and Oceanography, 52, 635-648, 2007.

Möllmann, C., Diekmann, C., Müller-Karulis, B., Kornilovs, G., Plikshs, M., and Axe, P.: Reorganization of a large marine ecosystem due to atmospheric and anthropogenic pressure: a discontinuous regime shift in the Central Baltic Sea. Global Change Biology, 15, 1377-1393, 2009.

Newman, M., Compo, G. P., and Alexander, M. A.: ENSO-Forced Variability of the Pacific Decadal Oscillation. Journal of Climate, 16, 3853-3857, 2003.

Overland, J., Alheit, J., Bakun, A., Hurrell, J. H., Mackas, D. L., and Miller, A. J.: Climate controls on marine ecosystems and fish populations. Journal of Marine Systems, 79, 305-315, 2010.

Overland, J., Rodionov, S., Minobe, S., and Bond, N.: North Pacific regime shifts: Definitions, issues and recent transitions. Progress in Oceanography, 77, 92-102, 2008.

Page, E. S.: Continuous Inspection Scheme. Biometrika, 41, 100–115, 1954.

Palmer, J. and Totterdell, I.: Production and export in a global ocean ecosystem model. Deep Sea Research Part I: Oceanographic Research Papers, 48, 1169-1198, 2001.

Paternoster, R., Brame, R., Mazerolle, P., and Piquero, A.: Using the correct statistical test for the equality of regression coefficients. Criminology, 36, 859-866, 1998.

Polovina, J. J., Mitchum, G. T., and Evans, G. T.: Decadal and basin-scale variation in mixed layer depth and the impact on biological production in the Central and North Pacific, 1960-88. Deep Sea Research Part I: Oceanographic Research Papers, 42, 1701-1716, 1995.

Reid, P. C., Borges, M. F., and Svendsen, E.: A regime shift in the North Sea circa 1988 linked to changes in the North Sea horse mackerel fishery. Fisheries Research, 50, 163-171, 2001.

Rodionov, S. N.: A sequential algorithm for testing climate regime shifts. Geophysical Research Letters, 31, L09204, 2004.

Rudnick, D. L. and Davis, R. E.: Red noise and regime shifts, Deep Sea Research Part I: Oceanographic Research Papers, 50, 691-699, 2003.

Scheffer, M., Bascompte, J., Brock, W., Brovkin, V., Carpenter, S. R., Dakos, V., Held, H., van Nes, E. H., Rietkerk, M., and Sugihara, G.: Early-warning signals for critical transitions. Nature, 461, 53-59, 2009.

Schwing, F. B.: Regime Shifts, Physical Forcing, in Encyclopedia of Ocean Sciences (Second Edition), edited by J. H. Steele, pp. 709-716, Academic Press, Oxford, 2009.

Smith, T. M. and Reynolds, R. W.: Extended reconstruction of global sea surface temperatures based on COADS data (1854–1997). Journal of Climate, 16, 1495-1510, 2003.

Smith, T. M., Reynolds, R. W., Peterson, T. C., and Lawrimore, J.: Improvements to NOAA's Historical Merged Land–Ocean Surface Temperature Analysis (1880–2006). Journal of Climate, 21, 2283-2296, 2008.

Spencer, M., Birchenough, S. N. R., Mieszkowska, N., Robinson, L. A., Simpson, S. D., Burrows, M. T., Capasso, E., Cleall-Harding, P., Crummy, J., Duck, C., Eloire, D., Frost, M., Hall, A. J., Hawkins, S. J., Johns, D. G., Sims, T. D., Smyth, T. J., and Frid, C. L. J.: Temporal change in UK marine communities: trends or regime shifts? Marine ecology, 32, 10-24, 2011.

Wolter, K. and Timlin, M. S.: Measuring the strength of ENSO events – how does 1997/98 rank? Weather, 53, 315-324, 1998.

Wunsch, C.: The interpretation of short climate records, with comments on the North Atlantic and Southern Oscillations. Bulletin of the American Meteorological Society, 80, 245-255, 1999.

Yatsu, A., Aydin, K. Y., King, J. R., McFarlana, G. A., Chiba, S., Tadokoro, K., Kaeriyama, M., and Watanabe, Y.: Elucidating dynamic responses of North Pacific fish populations to climatic forcing: Influence of life-history strategy. Progress in Oceanography, 77, 252-268, 2008.

Yool, A., Popova, E. E., and Anderson, T. R.: Medusa-1.0: a new intermediate complexity plankton ecosystem model for the global domain. Geoscientific Model Development, 4, 381-417, 2011.

Yool, A., Popova, E. E., and Anderson, T. R.: MEDUSA-2.0: an intermediate complexity biogeochemical model of the marine carbon cycle for climate change and ocean acidification studies. Geoscientific Model Development, 6, 1767-1811, 2013.

Table 1. Nutrient cycles and Plankton Functional Types represented in each model.

		HadOCC	Diat-HadOCC	MEDUSA	PlankTOM10	ERSEM
Nutrients	Nitrogen	x	x	x	x	x
	Phosphorous				x	x
	Silica		x	x	x	x
	Iron		x	x	x	x
	Carbon	x	x	x	x	x
	Alkalinity	x	x	x	x	x
Plankton Functional Type	Generic phytoplankton	x	x		x	
	Diatoms		x	x	x	x
	Large phytoplankton					x
	Picoplankton			x	x	x
	Coccolithophores				x	
	N_2 fixers				x	
	Flagellates					x
	Phaeocystis				x	
	Generic zooplankton	x	x			
	Microzooplankton			x	x	x
	Mesozooplankton			x	x	x
	Macrozooplankton				x	
	Heterotrophic nanoflagellates					x
	Bacteria				x	x
	Tracers	7	13	15	39	57

Table 2. List of models fitted in this study with their associated Schwarz Information Criterion (SIC) formulation.

Model description	Equations
(I) Constant mean	$y_t = \mu + \varepsilon_t \quad (t = 1,...,n)$ where y_t represents the time series, μ is the mean, ε_t are the random errors, t is the time and n is the length of the time series $SIC_I = n\log(RSS) + n(1+\log(2\pi)) + (2-n)\log(n)$ $RSS = \sum_{t=1}^{n}(y_t - \hat{\mu})^2$, where $\hat{\mu}$ is the maximum likelihood estimates of μ
(II) Shift in the mean	$y_t = \begin{cases} \mu_1 + \varepsilon_t & (t = 1,...,p) \\ \mu_2 + \varepsilon_t & (t = p+1,...,n) \end{cases}$ where μ_1 and μ_2 are the means before and after the shift at time p $SIC_{II}(p) = n\log(RSS) + n(1+\log(2\pi)) + (3-n)\log(n)$ $RSS = \sum_{t=1}^{p}(y_t - \hat{\mu}_1)^2 + \sum_{t=p+1}^{n}(y_t - \hat{\mu}_2)^2$, where $\hat{\mu}_1$ and $\hat{\mu}_2$ are the maximum likelihood estimates of μ_1 and μ_2
(III) Linear trend	$y_t = \lambda + \beta t + \varepsilon_t \quad (t = 1,...,n)$ where λ is the intercept and β the trend of the linear regression model $SIC_{III} = n\log(RSS) + n(1+\log(2\pi)) + (3-n)\log(n)$ $RSS = \sum_{t=1}^{n}(y_t - \hat{\lambda} - \hat{\beta}t)^2$, where $\hat{\lambda}$ and $\hat{\beta}$ are the maximum likelihood estimates of λ and β
(IV) Shift in the intercept and same linear trend	$y_t = \begin{cases} \lambda_1 + \beta t + \varepsilon_t & (t = 1,...,p) \\ \lambda_2 + \beta t + \varepsilon_t & (t = p+1,...,n) \end{cases}$ where λ_1 and λ_2 are the intercept before and after the shift $SIC_{IV}(p) = n\log(RSS) + n(1+\log(2\pi)) + (4-n)\log(n)$ $RSS = \sum_{t=1}^{p}(y_t - \hat{\lambda}_1 - \hat{\beta}t)^2 + \sum_{t=p+1}^{n}(y_t - \hat{\lambda}_2 - \hat{\beta}t)^2$, where $\hat{\lambda}_1$, $\hat{\lambda}_2$ and $\hat{\beta}$ are the maximum likelihood estimates of λ_1, λ_2 and β
(V) Shift in both the intercept and linear trend	$y_t = \begin{cases} \lambda_1 + \beta_1 t + \varepsilon_t & (t = 1,...,p) \\ \lambda_2 + \beta_2 t + \varepsilon_t & (t = p+1,...,n) \end{cases}$ where β_1 and β_2 are the trend before and after the shift $SIC_V(p) = n\log(RSS) + n(1+\log(2\pi)) + (5-n)\log(n)$ $RSS = \sum_{t=1}^{p}(y_t - \hat{\lambda}_1 - \hat{\beta}_1 t)^2 + \sum_{t=p+1}^{n}(y_t - \hat{\lambda}_2 - \hat{\beta}_2 t)^2$, where $\hat{\lambda}_1$, $\hat{\lambda}_2$, $\hat{\beta}_1$ and $\hat{\beta}_2$ are the maximum likelihood estimated of λ_1, λ_2, β_1 and β_2

* All these models rely on the assumption that the random errors are independent and identically normally distributed $\varepsilon_t \sim N(0,\sigma^2)$

Table 3. Results from change-point detection analysis on the first principal component (PC1) of each model. Years in bold have a significant shift (p-value < 0.05).

Model	Shift year	Shift type	SIC	SIC (Null model)	p-value
HadOCC	**1977**	mean	180.62	225.69	<0.01 [a, c]
DiatHadOCC	**1976**	mean	185.55	240.42	<0.01 [c]
MEDUSA	**1978**	trend and intercept	184.54	207.00	<0.01 [c]
PlankTOM10	1987	intercept	141.14	152.90	0.21 [c]
ERSEM	1987	intercept	189.09	192.55	0.63 [c]

[a] residuals not normally distributed (Lilliefors test, 5% critical level)

[b] residual variance not constant (Fisher test, 5% critical level)

[c] residuals not independent (Durbin-Watson test, 5% critical level): the Monte Carlo simulations to estimate the p-value incorporates the first-order autocorrelation of the residuals.

Table 4. Results of the principal component analysis: percentage of variance explained by the first principal component (PC1) and relative contributions of the different variables to this component.

Model	Variance explained (%)	Relative contribution (%)								
		SST	MLD	CHL	PP	PHY	ZOO	DIN	FE	SI
HadOCC	61.09	13.53	2.69	19.02	11.08	18.91	16.82	17.94	-	-
DiatHadOCC	63.42	10.41	1.92	13.88	12.84	13.93	13.51	11.85	12.50	9.16
MEDUSA	36.33	10.16	0.94	15.91	7.34	9.90	6.18	19.07	16.31	14.18
PlankTOM10	66.05	7.74	6.25	14.29	13.99	14.59	14.27	14.30	1.10	13.48
ERSEM	50.74	8.88	1.31	14.08	15.05	8.32	12.71	9.63	15.86	14.18

Table 5. Forcing-response regressions in HadOCC, Diat-HadOCC and MEDUSA with sea surface temperature (SST) as the physical forcing and surface chlorophyll (CHL), integrated primary production (PP), total surface phytoplankton (PHY) and zooplankton biomass (ZOO) as the responses. The slopes of the linear regressions between the forcing and response before and after the shift are compared using a test of equality of two regression slopes. Bold indicates significant slope differences (p-value < 0.05).

HadOCC

Forcing	Response	Slope 1957-1976 (standard error)	Slope 1977-2007 (standard error)	Test statistic	p-value
SST	CHL	-0.025 (0.028)	-0.008 (0.024)	1.407	0.166
	PP	0.000 (0.005)	0.021 (0.011)	-1.703	0.095
	PHY	-0.008 (0.014)	-0.030 (0.013)	1.179	0.245
	ZOO	0.002 (0.004)	-0.012 (0.003)	2.823	**0.007**

Diat-HadOCC

Forcing	Response	Slope 1957-1976 (standard error)	Slope 1977-2007 (standard error)	Test statistic	p-value
SST	CHL	-0.121 (0.071)	-0.217 (0.052)	1.095	0.279
	PP	-0.033 (0.012) [b]	-0.022 (0.012)	-0.666	0.508
	PHY	-0.028 (0.025)	-0.069 (0.018)	1.345	0.185
	ZOO	-0.002 (0.006)	-0.018 (0.005)	2.034	**0.048**

MEDUSA

Forcing	Response	Slope 1957-1976 (standard error)	Slope 1977-2007 (standard error)	Test statistic	p-value
SST	CHL	0.002 (0.006)	-0.013 (0.007)	1.476	0.146
	PP	0.019 (0.004)	0.020 (0.005)	-0.129	0.898
	PHY	0.014 (0.004)	0.006 (0.004)	1.458	0.151
	ZOO	0.039 (0.007)	0.027 (0.007) [b]	1.132	0.263

[a] residuals not normally distributed (Lilliefors test, 5% critical level)

[b] residual variance not constant (Breusch Pagan test, 5% critical level)

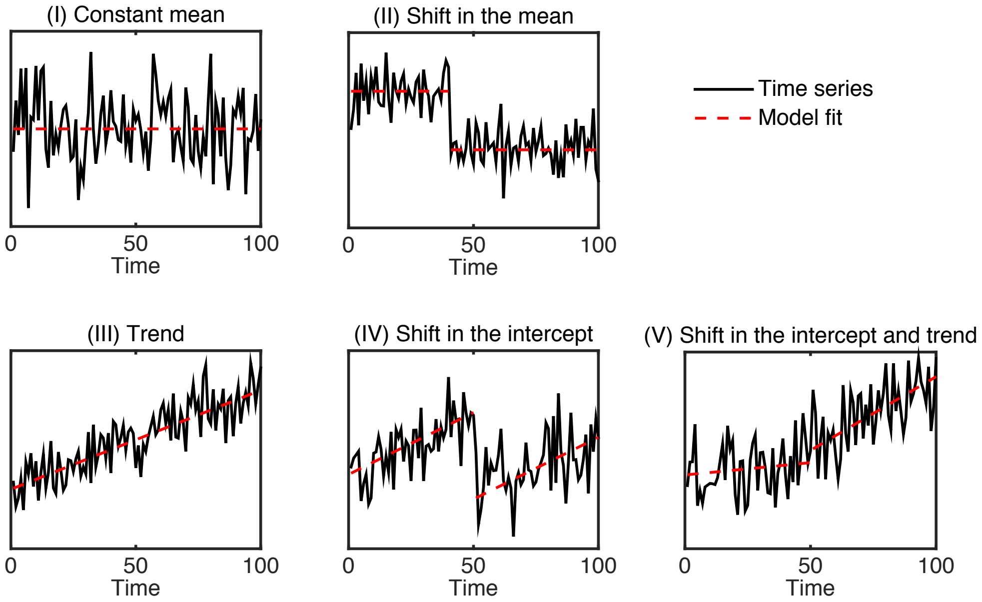

Figure 1. Five types of statistical models that were fitted to the data. The solid lines are synthetic time series drawn from a model with (I) a constant mean, (II) shift in the mean, (III) trend, (IV) shift in the intercept of the trend (the trend is the same before and after the shift) and (V) shift in both the intercept and trend. The constant mean (I) is the null model for a shift in the mean (II) when testing for significance. Similarly, the trend model (III) is the null model to test the shift significance when the model selected is either a shift in the intercept (IV) or a shift in both the intercept and trend (V). The corresponding models are further described in Table 2. Figure adapted from Beaulieu et al. (2012).

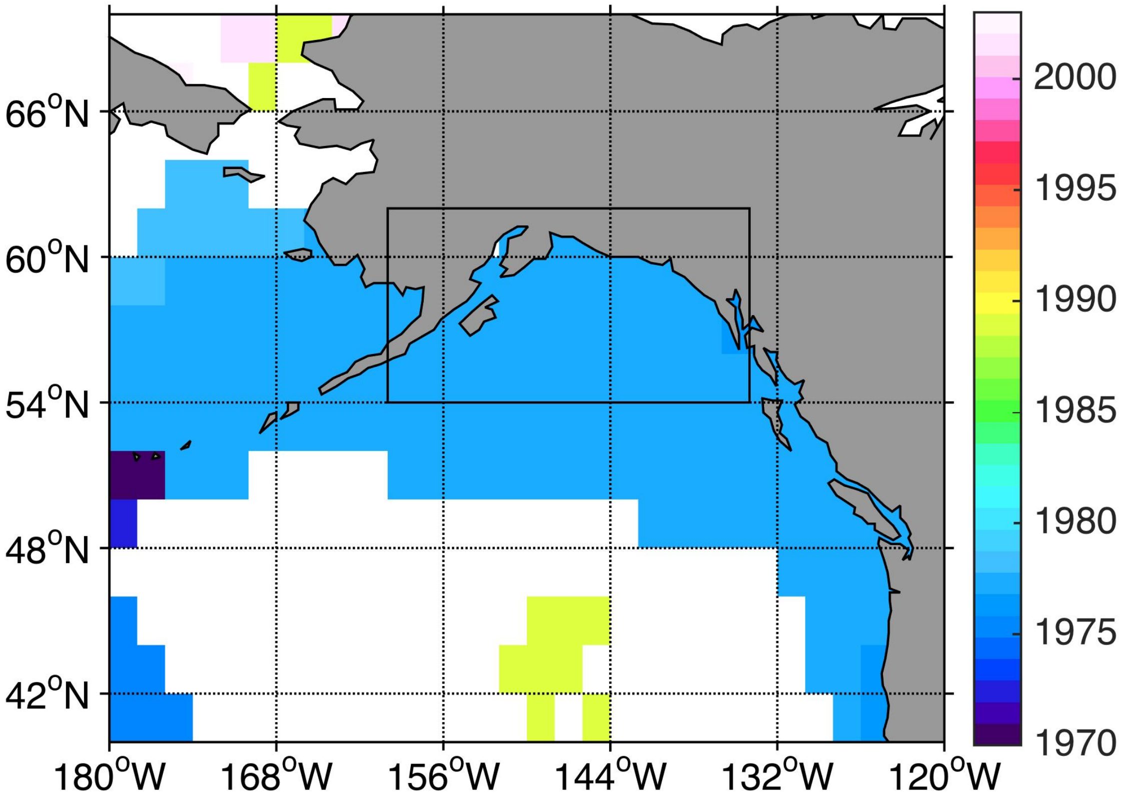

Figure 2. Timing of shift detected in observed sea surface temperature in the North Pacific using change-point analysis showing a predominant signal in 1977. White areas indicate where a shift is not significant (p-value < 0.05). The black box indicates the Gulf of Alaska region used in this study.

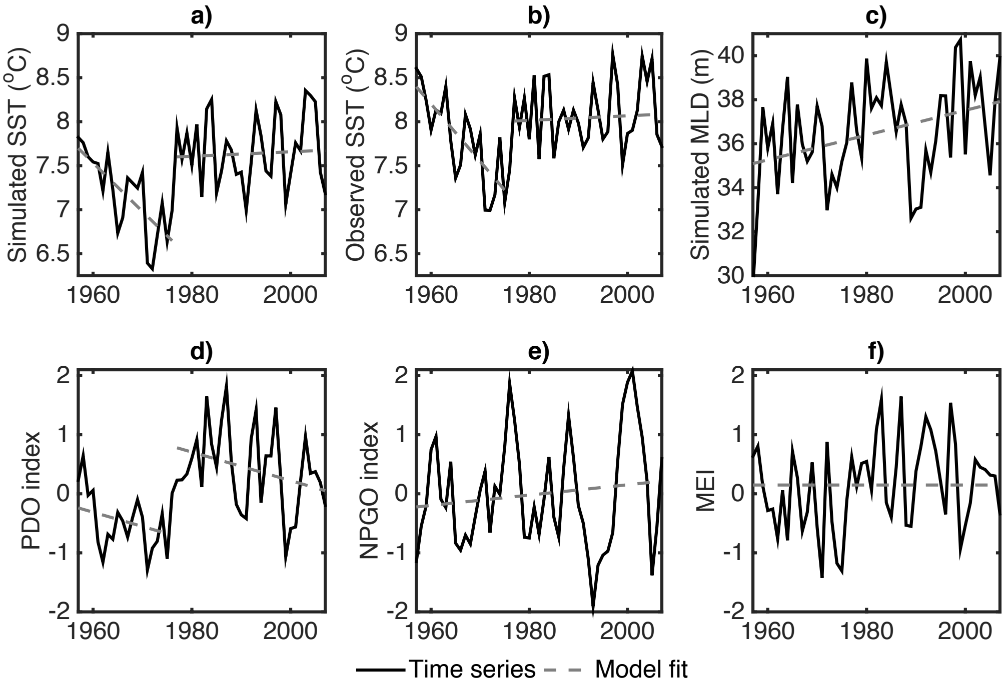

Figure 3. Time series of (a) simulated sea surface temperature (SST), (b) observed SST and (c) simulated mixed layer depth (MLD) for the Gulf of Alaska. The simulated time series of SST and MLD are the same in the five ocean models used. Time series of large-scale oscillations representing the climate in the Gulf of Alaska: (d) Pacific Decadal Oscillation (PDO) index, (e) North Pacific Gyre Oscillation (NPGO) index and (f) Multivariate El Niño Southern Oscillation index (MEI). The grey dotted lines represent the statistical model chosen (see Table A1) to fit these time series. Both the simulated SST and observed SST exhibit a significant shift in intercept and trend occurring in 1976 (p-value < 0.05, see Table A1). The MLD time series does not exhibit a significant shift and is best represented by a linear trend. Among the large-scale oscillations, only the PDO exhibits a significant shift in 1976.

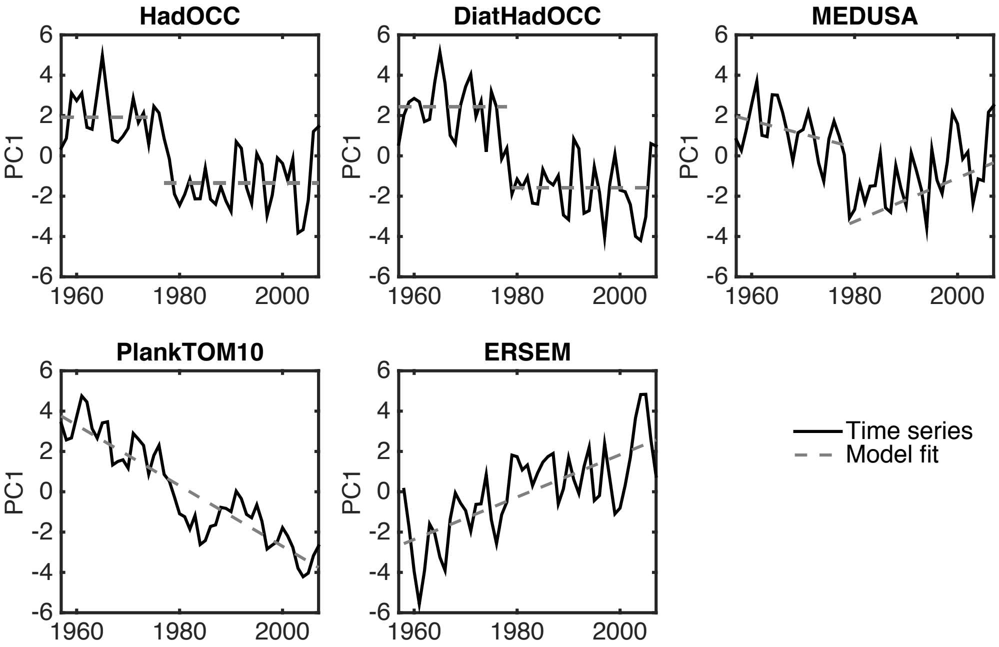

Figure 4. First principal component (PC1) of sea surface temperature, mixed layer depth, surface dissolved inorganic nitrogen, silica, iron, surface chlorophyll, integrated primary production, total surface phytoplankton and zooplankton biomass (if available) averaged over the Gulf of Alaska for each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

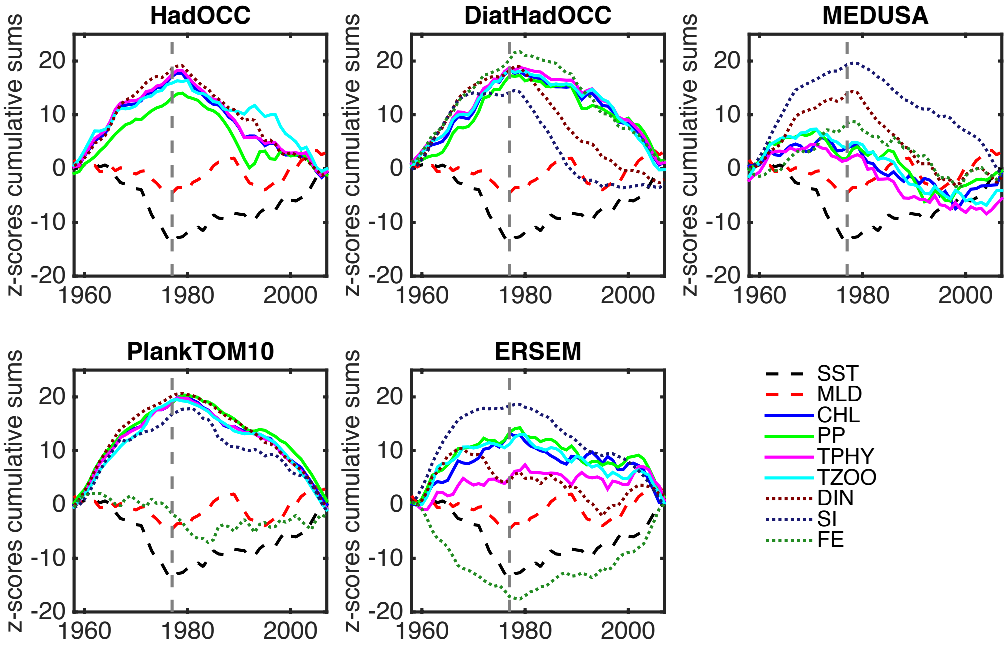

Figure 5. Cumulative sums of the z-scores of simulated sea surface temperature (SST), mixed layer depth (MLD), surface dissolved inorganic nitrogen (DIN), silica (SI), iron (FE), surface chlorophyll (CHL), integrated primary production (PP), total surface phytoplankton (PHY) and zooplankton (ZOO) biomass for each model averaged over the Gulf of Alaska region. Z-scores are calculated by subtracting the mean and dividing by the standard deviation of each time series. Cumulative sums of the z-scores are then calculated. The vertical lines in 1977 provide a guide to the eye showing where the slopes change after 1977.

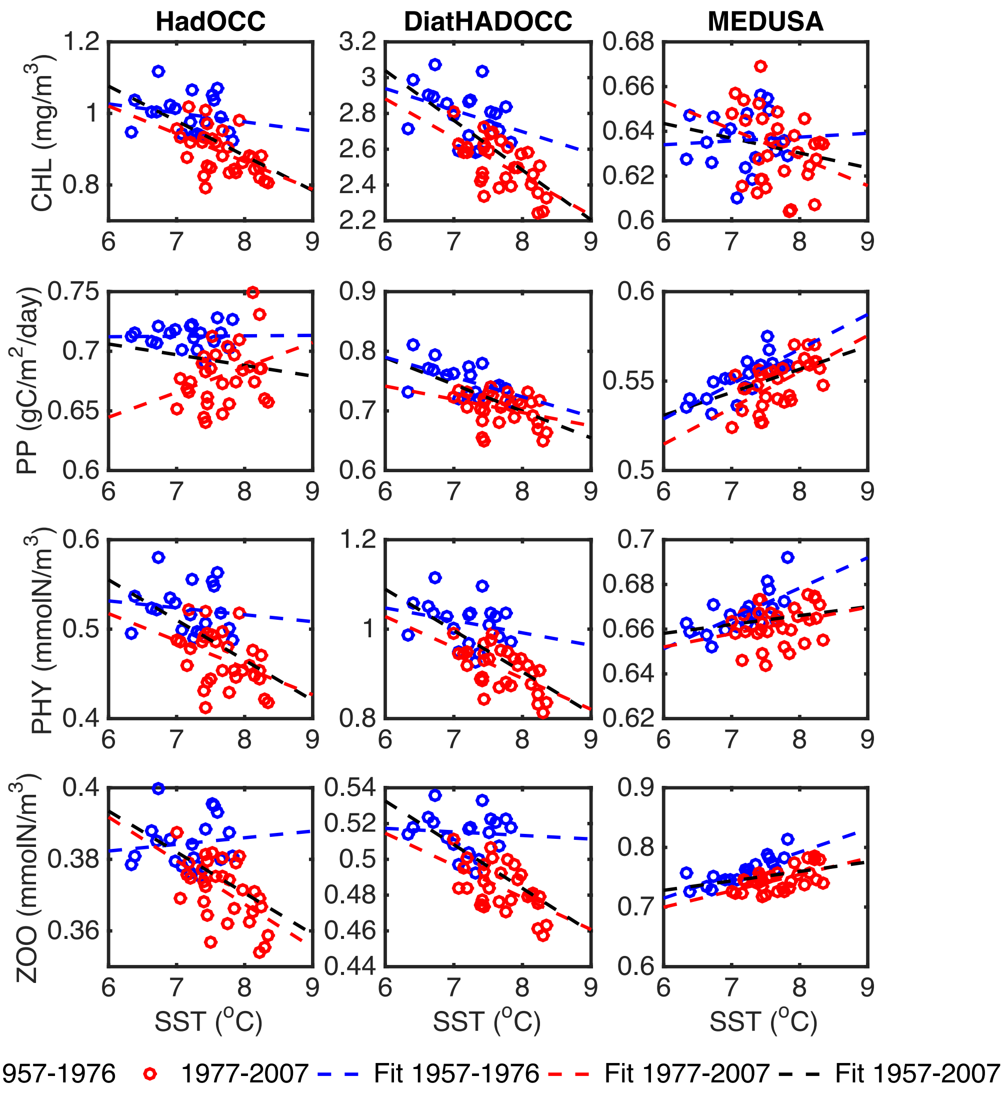

Figure 6. Relationships matrix between simulated sea surface temperature (SST) and the biological variables over the Gulf of Alaska region. Columns represent different models (HadOCC, DiatHadOCC and MEDUSA) and rows represent different biological variables (surface chlorophyll (CHL), integrated primary 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.