Response to Reviewers by O. Andrews

We thank the reviewers for their helpful and constructive comments which have significantly improved the manuscript. We have made the following general modifications to address the reviewer's main points:

- **Added new Figure 2 that compares observed and simulated zonal mean [O_2] distributions, and included with this a more in-depth discussion of model performance.**
- **Added new Table 2 that compares two estimates of internal variability in the observations (from long-term [O_2] time series measurements) with model simulated internal variability.**
- **Significantly expanded the text with more discussion of statistical techniques and parameter choices, along with revisions for clarity.**

We hope that our revised manuscript will satisfy the comments of the reviewers. Please find below our detailed responses (emboldened), which follow each reviewer comment.

Response to Anonymous Referee #1 (RC C5499)

The present study investigates the detectability of externally forced oceanic O2 changes with respect to natural variability. The authors use an optimal fingerprinting method and two CMIP5 Earth system models to determine the cause of changes in the marine O2 distribution as recorded by WOCE and earlier O2 data. Overall evaluation: The topic of the study is extremely relevant for a broad scope of readers from climate modellers to conservationists. However, I do have a major point of criticism regarding the degree of uncertainty introduced by the lack of knowledge about the natural background variability of O2 and the way this is (mainly not) discussed in the paper. Apart from that the paper is well written and I am confident that it will be publishable after my claims have been addressed.

As detailed below, we have expanded the text to better discuss our estimate of natural internal variability in [O_2] and the measures taken to ensure that this estimate is robust.

Major point of criticism: After reading the paper my impression is that the authors are using their undeniable statistical skills in a somehow selective manner. On the one hand they are using truncated EOFs and an Optimal Fingerprinting method, on the other hand they do very little to constrain the crucial parameters v(0) and v(i) and fail to discuss the uncertainty associated with their choice of simply using the simulated O2 variability as a reference.

Constraints on v_i (signal error): We apply a standard method used to account for signal error (v_i) in climate change optimal detection by using TLS regression (e.g. Allen and Stott, 2003). This is a variant of linear regression widely employed in optimal fingerprinting studies, which provides a more conservative estimate of scaling factors because it includes noise in the model fingerprints. Signal error is thus already included in the method, and the only parameter that we specify is the amount of noise found in the models (x_i) relative to that which is found in the

observations (determined by the ensemble size of the CMIP5 experiment [see Table 1]). The description of this choice of fitting technique along with the inverse relationship between historical model ensemble size (*n*) and signal error has been better explained in the text of Section 2.2 (along with an added citation which describes the application of TLS regression to detection and attribution using GCMs [Stott et al., 2003]):

"Signal error (v_i) is inversely proportional to the model ensemble size (*n*) and can negatively bias scaling factors (Allen and Tett, 1999), particularly for variables where the forced response is small relative to internal variability. Thus, TLS is widely used in optimal detection studies as a more conservative approach which explicitly accounts for the effect of noise in simulated response patterns relative to that which is found in the observations (e.g. Stott et al., 2003; Terray et al., 2012; Stott et al., 2008)."

Constraints on v_0 (noise in the observations): Although we estimate internal variability in the observations from ESMs our method checks that modelled variability provides an adequate representation of observed climate noise for our purposes. The choice to derive internal variability in the observations (v_0) from control integrations of ESMs is based on the underlying rationale of optimal fingerprinting that internal variability is not easily estimated from observations on large spatial and temporal scales, and that piControl data provides a pure expression of the variability of the internal system. Moreover, observations of [O_2] contain signals driven by solar, volcanic and anthropogenic forcings. This point has been made more clearly in revised Section 2.2.

Considerable effort has been taken in our study to provide a reliable estimate of internal variability in [O_2] (v_0). We employ the standard residual consistency *F*-test which guides our analysis to retain a reduced dimension space in which internal variability is well simulated by models. This technique provides an effective strategy for evaluating whether the residual observed variance is consistent with model simulated internal variability at a given confidence level. Since the residual test passes for all experiments at our chosen truncation (apart from the Atlantic basin which marginally fails) we have confidence that piControl data has power as an independent estimate of noise in the observations. We have further explained this test in the Methods Section 2.2:

"In order to avoid spurious detection it is a necessary prerequisite that the internal variability estimated from control simulations (v_0) provides a realistic estimate observed climate noise in [O_2]. As such, the number of EOFs retained in the optimal fingerprint analysis is guided by checking the fidelity of model simulated internal variability against the residual observed variance at *k* truncations using a standard residual consistency *F*-test (Allen and Tett, 1999). This check is used to test the null hypothesis that internal variability as simulated by models is consistent with observed variability on the scales retained in the analysis."

The authors claim that "The instrumental record of dissolved O2 measurements is not sufficiently long to get a reliable approximation of internal climate variability (v0), and also includes perturbations driven by external forcing." However, surface O2 data are available for HOT (1989-2012), BATS (1988-2012) and ESTOC (1994-2012). Of course this is not really sufficient and it includes perturbations, but it could give us at least some idea whether or not the two ESMs do a reasonable job in reproducing the interannual O2 variability. What would an underestimation of the internal variability (as mentioned in the Discussion) mean for the potential detectability ant the present conclusions of the paper? A signal-to-ratio study has to come up with a reliable estimate for the noise. If the available data cannot provide such an estimate we cannot make a reliable statement about the signal detectability.

As suggested by Reviewer 1 we have further justified our choice of v_0 by comparing internal variability estimated from piControl data with available [O_2] time series data from Ocean Station Papa (1956–2007, Eastern North Pacific; Whitney et al., [2007]) and the Oyashio Current region (1968–1998, Western Subarctic Pacific; Ono et al. [2001]). This is presented in new Table 2 (see below) with a description included in Section 2.2:

"In addition to the residual consistency test, we assess the reliability of model simulated climate variability by comparing piControl output with detrended subsurface [O_2] measurements from two long-term time series: Ocean Station Papa (1956–2007, Whitney et al., 2007) and the Oyashio Current region (1968–1998, Ono et al. 2001). Observed decadal standard deviations calculated for both time series fall within the 10–90% ranges of MPI-ESM-LR control simulation estimates, demonstrating that this model provides a robust estimate of internal variability in [O_2] on decadal timescales (Table 2). The HadGEM2-ES control simulation significantly underestimates decadal variability in [O_2] when compared to time series data and is thus less reliable than simulations from MPI-ESM-LR in the context of our analysis. "

Table 2. Comparison of decadal standard deviations of [O_2] (μmol kg^{-1}) between observations and CMIP5 piControl experiments at Ocean Station Papa in the Eastern North Pacific (50°N, 145°W, σ = 26.9 [~ 250 – 350 m]) and the Oyashio Current region in the subarctic Western North Pacific (39° – 42°N, 143° – 145°E, σ = 26.9 [~ 400 – 450 m]). piControl output from MPI-ESM-LR and HadGEM2-ES is sampled to calculate multiple estimates of decadal standard deviations for each time series. The mean and (10th – 90th) percentile ranges of piControl estimates are shown. Observations are detrended using a linear fit to extract variability. Drift has been removed from the control integration of HadGEM2-ES as described in Section 2.1.

	Observations	MPI-ESM-LR	HadGEM2-ES
Ocean Station Papa	11.1	9.1 (6.2 – 12.0)	5.3 (3.9 – 6.7)
Oyashio Current region	8.1	9.9 (4.9 – 15.1)	1.4 (0.4 – 2.4)

Finally, in order to avoid confusion, we have augmented the paragraph in Section 4 (Discussion) which explores the implications of models underestimating variability in observed [O_2]. Whilst this is an interesting topic for discussion such a statement could wrongly imply that v_0 is demonstrably underestimated in our analysis, when the intention was more to generally introduce the topic of simulated climate variability in ESMs and ocean-only hindcasts. To address this

we include a statement in Section 4 to clarify the steps taken to ensure a reasonable estimate of internal climate noise:

"However, the piControl derived estimates of natural internal variability used in this analysis are shown to be consistent with observed variance using a standard residual consistency test and comparison of simulated noise with two observational estimates (Sect. 2.2 and Table 2)."

Minor points:

p. 12471 / l. 16 A Corrigendum has been published for the Schmittner et al. [2008] paper reporting an error in the calculation of light limitation. The authors might want to check if the results of the paper can be cited in this context.

We have checked the corrigendum and there are no implications for our study. Centennial scale increases in the volume of suboxia as reported by Schmittner et al. (2008) are a robust result and have been cited elsewhere since the corrigendum (e.g. Duteil and Oschlies, 2011; Peña et al., 2010).

p 12474 / l. 1 What is meant by "is taken into account in this data using an a priori noise estimate, ..."?

This statement has been reworked and refers to the approach taken to characterise natural variability in [O_2] data a priori (Helm et al., 2011; Bindoff and Wunsch, 1992. We have improved the explanation of this in the text. New text:

"Noise variance calculated using the difference between neighbouring data points is used to provide an a priori estimate of natural variability in [O_2] data (Helm et al., 2010, Helm et al., 2011; Bindoff and Wunsch, 1992). This technique accounts for mesoscale processes and to some extent longer period internal variability such as the dominant climate modes."

p. 12487 / l. 8 It is not really "our understanding of the background internal variability" but the authors choice to use the simulated O2 variability as a reference.

Accepted and changed to "internal variability as simulated by models".

Figure 5: y-label should read "Change in ..."

Accepted and axis label corrected to read "Change in Stratification Index".

Figure 6: On the one hand the authors make the effort to derive the reference variability shading from the "diagonal of the autocovariance matrices estimated by sampling model piControl simulations" but they do not discuss the fact that a simple factor 2 in the estimate for the background O2 variability would result in very different conclusions regarding the detectability of O2 changes.

As stated above we validate our choice of model simulated internal variability using (1) a residual consistency check; and in the revised manuscript (2) via

comparison with [O_2] time series data from two locations. The internal variability of the models is represented statistically through v_0 (as discussed above). This covariance matrix takes into account the covariance with neighbouring spatial points. Consequently the diagonal elements plotted on Figure 7 (previously Figure 6) provide a simple diagnostic which illustrates the latitudinal variations in internal variability but do not show the full relationships of the covariance matrices and its consequences on the estimates of scaling factors.

Looking at Figures 2-5, the models do a rather poor job in reproducing the observed changes. This is not the authors fault but some of the discussions of the model-data comparisons sound more like wishful thinking.

A thorough discussion of model-data agreement has been provided in response to Reviewer 2 (see Response to Anonymous Referee #2 below). Section 3.1 (Model-data comparison) has also been edited.

Response to Anonymous Referee #2 (RC C5529)

This paper applies a statistical technique to the output of two state-of-the-art climate models participating in the IPCC 5th Assessment Report (the CMIP5 model suite) to investigate whether observed changes in ocean oxygen content can be explained by anthropogenic forcing or are simply due to natural variability. Using an optimal fingerprinting method they find that the changes are inconsistent with internal variability (as simulated in the climate models). Specifically, these results are robust for depth and zonal-mean O2 changes for the global ocean and for zonal-mean changes for the global ocean and Pacific basin. Changes in the Atlantic basin are found to be consistent with natural variability.

The paper addresses an important question in climate science. Much research suggests that deoxygenation will occur as the climate warms, and while observations do in fact suggest this, it has remained unclear whether the changes taking place are natural or anthropogenic.

This is a paper that’s definitely appropriate for BG but I have concerns. Primary among them is the fact that Figs. 2 and 3 comparing model and data don’t inspire much confidence in even the most state-of-the-art climate models. Not only is the magnitude of the changes severely underestimated by the models, in some cases even the sign is incorrect. These aren’t point-wise comparisons where one would hardly expect the models to perfectly replicate reality. Instead, these are plots of very large scale features. It is unclear to me whether it makes any sense to apply sophisticated statistical techniques to tease out small signals to such models at all.

Indeed the reviewer is right that there is always an unavoidable structural uncertainty in optimal fingerprinting studies (e.g. Hegerl and Zwiers, 2011) driven by the representation of physical and biogeochemical processes in models, which can cause systematic errors in model response patterns. The use of several models and several spatial averaging schemes allows us to generate multiple model fingerprints to better quantify the effects of possible structural

errors in model response patterns. We now include the following statement in Section 2.2:

"The use of several models and several spatial averaging schemes provides multiple model fingerprints that are used to quantify possible errors in model response patterns ("structural uncertainty") driven by inadequate representation of physical and biogeochemical process in ESMs (e.g. Hegerl and Zwiers, 2011)."

However, as we show in our study, positive detection results ($\beta > 0$) for global and Pacific basin detection experiments demonstrate that significant correlations exist between model simulated and observed [O_2] changes. This is reinforced by the global 2-D detection experiment where 5–95% uncertainty bounds on β are consistent with 1, passing the "attribution test" of model-data consistency in amplitude. Moreover, if the models were insufficient to detect trends in [O_2] due to structural error statistical tests performed in our study would fail. Specifically, the residual consistency *F*-test provides information about model-data agreement and structural error in the optimal detection analysis. Where the residual test passes (for all global experiments and the Pacific basin individually) we have some confidence that, in a statistical sense, the residual of the TLS regression agrees with model estimates of noise, as would be expected if models have capacity to estimate the forced signal and that this response adequately fits the observations. Text has been added to Sections 2.2 and 3.2 to explain that the residual consistency test is also used to diagnose possible inconsistencies between forced model responses and observations:

Section 2.2 "Failure of the residual consistency test could also indicate that the timing or pattern of ESM response is incorrect"

Section 3.2: "The residual consistency test passes for both model experiments indicating no inconsistency between residual observed variance and model simulated internal variability, and suggesting that both ESMs simulate the externally forced signal adequately to explain observed [O_2] changes."

The statistical tests succeed because generally models either reproduce the sign of zonal mean [O_2] change (Fig. 3) or where they don't their response falls within the envelope of internal variability, as shown in Figure 7. Significant model-data disagreement in the 1-D case is evident at ~ 60°S, and is explained in Section 4 as a potential limitation of this study likely driven by elevated ventilation in the Southern Ocean. We are also able to better investigate the degree of model-data agreement by using 2-D fingerprints, which demonstrate that models have significant capacity to reproduce the main features of [O_2] change as a function of latitude and depth (Fig. 4 and Fig. 5). MPI-ESM-LR in particular has skill in simulating a pattern of high latitude deoxygenation throughout the water column countered by a region of [O_2] increase within the ocean interior at low latitude. The signal-to-noise ratios of observations and model response patterns are also maximised via normalisation by internal variability (v_0).

The TLS detection method assumes that structural uncertainties in models are driven by internal variability. This assumption holds in our study, with errors in model response patterns being most prominent at mid–high latitudes where

climate noise is largest, and generally falling within the spread of internal variability (as shown in Figure 7). We have clarified this in Sect. 3.2 and Sect. 4:

Section 3.2: “Structural errors in the pattern of [O_2] change simulated by both models also generally fall within the spread of internal variability.”

Section 4: “Qualitative model-data comparison (Sect. 3.1) suggests that regional differences between model simulated and observed patterns of [O_2] change could also contribute to the weaker simulated zonal mean signal. However, consistent with the assumption of the TLS detection model that “structural uncertainty” has the same structure as internal variability (e.g. Terray et al., 2012), model errors generally fall within the range of internal variability (Fig. 7).”

Finally, it is worth highlighting that whilst detection studies with a hypothetical ESM which reproduces the observations exactly would have the most power in detecting forcings in observed changes, optimal fingerprinting also provides an effective and rigorous strategy for evaluation of climate model simulations against observations. For example, *β* values greater than one indicate that there are missing forcings or errors in model responses, and failure of the residual consistency test in the Atlantic basin suggests deficiencies in model fingerprints or simulated variability (see discussion in Sect. 3.2 and Sect. 4). A statement about the utility of optimal fingerprinting in model evaluation has also been added to the introduction of Section 2.2 for context:

“This statistical technique is widely used in the detection and attribution of climate change (e.g. IDAG, 2005; Hegerl et al., 2010) and also provides a powerful test of ESM performance which accounts for the effect of natural internal variability.”

On a similar note, I would have liked to see how well the models do with respect to the O2 distribution itself. A figure comparing the simulated O2 distribution at different depths or vertical profiles with data would be very helpful.

We have added a figure showing zonal mean sections of observed and simulated (MPI-ESM-LR and HadGEM2-ES) [O_2] distributions (new Figure 2.). A comparison of model simulated [O_2] distribution with observations has also been included in Section 3.1. Generally the latitude-depth features of observed zonal mean [O_2] distribution are reproduced by both models. A comprehensive model-data evaluation of biogeochemical tracers in MPI-ESM-LR CMIP5 historical experiments at different depth levels has recently been undertaken by Ilyina et al. (2012). Reference to this paper is now also included in Section 3.1. Our presentation of zonal mean [O_2] sections complements the analysis of Ilyina et al. (2012), which plots horizontal [O_2] distributions at different depth levels. We added the following new text in Section 3.1:

“We evaluate model performance by comparing simulated and observed zonally averaged [O_2] distributions for the ~1992 time period. Historical integrations of MPI-ESM-LR and HadGEM2-ES are able to capture the general latitude-depth pattern of [O_2] distribution present in the observations (Fig. 2). The [O_2] minimum between 20°S and 20°N is reproduced at mid-depths by both models, with increasing [O_2]

towards higher latitudes in both hemispheres. However, HadGEM2-ES simulates higher than observed oxygen concentrations in this region. MPI-ESM-LR overestimates the spatial extent of low-latitude oxygen minimum waters and produces erroneously high [O_2] south of 60°S. Ilyina et al. (2012) present a detailed comparison between biogeochemical tracers in MPI-ESM-LR CMIP5 historical simulations and observations using a range of statistical metrics to assess model capability."

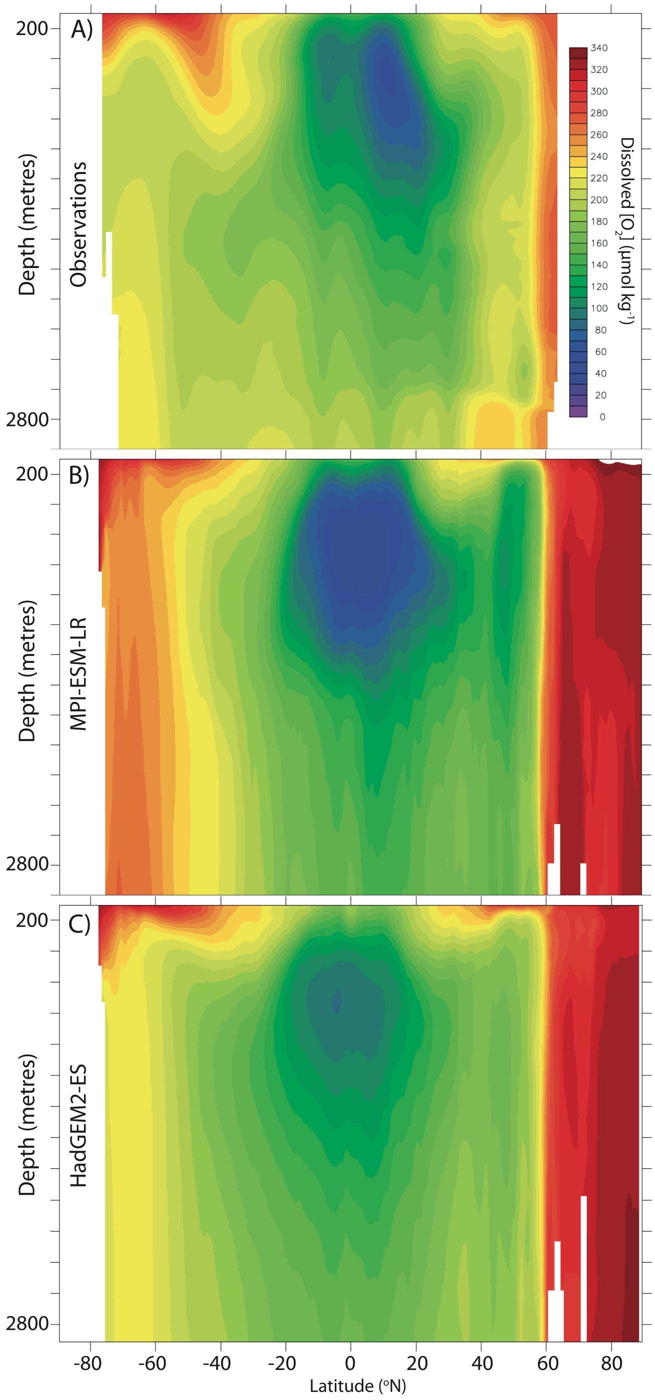

Fig. 2. Zonal mean [O_2] distribution (μmol kg^{-1}) for ~1992 as a function of latitude and depth from (A) observations (Helm et al., 2011), and historical integrations of (B) MPI-ESM-LR and (C) HadGEM2-ES.

Second, I found the description of the optimal fingerprinting method very hard to understand. A clearer, intuitive description of the method would be greatly beneficial to those of us who lack the requisite mathematical and statistical knowledge.

Section 2.2 has been extensively revised and edited for clarity.

Third, to the extent I understood their approach, the largest uncertainty seems to be in the selection of values of the noise parameters v_i and v_0. Models tend to underestimate variability and the ocean is typically undersampled so this is an issue with the observations as well. I found the discussion of this and the dependence of the results on their choice of these parameters inadequate.

A significant discussion and justification of v_i and v_0 has been undertaken in response to comments of Reviewer 1 (see Response to Anonymous Referee #1 above).

Incidentally, a simple time series of control and historical model simulations and observations at places where such data exist (the ocean time series stations?) would be very helpful here.

As detailed in the response to Reviewer 1 (see Response to Anonymous Referee #1 above) we have included new Table 2 which compares decadal standard deviations from two long ocean times series with MPI-ESM-LR and HadGEM2-ES piControl output sampled at these locations.

Also on the topic of their method, I was surprised by the sheer number of EOFs used (O(40)). I did not quite follow how this number is selected but in most other contexts I have seen the Akaike Information Criterion (AIC) used. Why not here?

We use the residual check (as is standard practise in optimal fingerprinting) to guide the choice of EOF truncation in our analysis towards a dimension space within which models simulate v_0 reasonably well (e.g. Allen and Stott, 2003). The common use of EOFs to describe the fundamental signal in data, in this case, tends to minimise the number of modes retained. However, for spatially resolved 2-D data more than 40 EOFs can be necessary to explain the variance in the data, particularly where higher wavenumber modes of internal variability become important (e.g. Terray et al., 2012). This point is made more clearly in revised Section 2.2.

Sec. 3.1 first sentence: "marked zonal" should be "marked meridional" I think.

Accepted and changed.

In summary, the authors tackle an interesting and important topic. On the one hand, the authors should be given credit for approaching the problem rigorously and making use of climate model simulations. On the other, what is one to make of results based on models that obviously perform very poorly even at the largest scales (by the particular metric of reproducing O2 changes). That said, on balance I am inclined to support eventual publication of the paper.

References

Allen, M. R. and Tett, S. F. B.: Checking for model consistency in optimal fingerprinting, Clim. Dynam., 15, 419–434, 1999.

Bindoff, N.L. and Wunsch, C.: Comparison of Synoptic and Climatologically Mapped Sections in the South Pacific Ocean, J. Climate, 5, 631–645, 1992.

Duteil, O. and Oschlies, A.: Sensitivity of simulated extent and future evolution of marine suboxia to mixing intensity, Geophys. Res. Lett., 38, L06607, doi:10.1029/2011GL046877, 2011.

Hegerl, G., & Zwiers, F.: Use of models in detection and attribution of climate change, WIREs Clim. Change, 2, 570–591, doi: 10.1002/wcc.121, 2011.

Helm, K. P., Bindoff, N. L., and Church, J. A.: Changes in the global hydrological-cycle inferred from ocean salinity, Geophys. Res. Lett., 37, L18701, doi:10.1029/2010GL044222, 2010.

Helm, K. P., Bindoff, N. L., and Church, J. A.: Observed decreases in oxygen content of the global ocean, Geophys. Res. Lett., 38, L23602, doi:10.1029/2011GL049513, 2011.

Ilyina T., Six, K. D., Segschneider, J., Maier-Reimer, E., Li, H., and Núñez-Riboni, I.: The global ocean biogeochemistry model HAMOCC: Model architecture and performance as component of the MPI-Earth System Model in different CMIP5 experimental realizations, J. Adv. Model. Earth Syst., submitted, 2012.

Ono, T., Midorikawa, T., Watanabe, Y. W., Tadokoro, K., and Saino, T.: Temporal increases of phosphate and apparent oxygen utilization in the subsurface waters of western subarctic Pacific from 1968 to 1998, Geophys. Res. Lett., 28, 3285–3288, 2001

Peña, M. A., Katsev, S., Oguz, T., and Gilbert, D.: Modeling dissolved oxygen dynamics and hypoxia, Biogeosciences, 7, 933–957, doi:10.5194/bg-7-933-2010, 2010.

Schmittner, A., Oschlies, A., Matthews, H. D., and Galbraith, E. D.: Future changes in climate, ocean circulation, ecosystems, and biogeochemical cycling simulated for a business as-usual CO2 emission scenario until year 4000 AD, Global Biogeochem. Cy., GB1013, doi:10.1029/2007GB002953, 2008.

Stott, P. A., Allen, M. R., and Jones, G. S.: Estimating signal amplitudes in optimal fingerprinting. Part II: application to general circulation models, Clim. Dynam., 21, 493–500, doi: 10.1007/s00382-003-0314-8

Stott, P. A., Sutton, R. T., and Smith, D. M.: Detection and attribution of Atlantic salinity changes, Geophys. Res. Lett., 35, L21702, doi:10.1029/2008GL035874, 2008.

Terray, L., Corre, S. Cravatte, T. Delcroix, Reverdin, G., and Ribes, A.: Near-surface salinity as Nature's rain gauge to detect human influence on the tropical water cycle, J. Clim., 25, 958–977, doi:10.1175/JCLI-D-10-05025.1, 2012.

Whitney, F. A, Freeland, H. J., and Robert, M.: Persistently declining oxygen levels in the interior waters of the eastern subarctic Pacific, Prog. Oceanogr., 75, 179–199, 2007.