**Responses to reviewers**

We would like to thank the reviewers for taking the time to read our manuscript and for their comments. Our responses to each review (font size 12), together with the original comments (font size 8), can be found below. Text quoted either from the old or the modified manuscript is in quotes.

## Response to reviewer #1

**Major comments:**

The authors neglect analytical errors in their analysis; "To isolate the effect of regression model selection from other sources of error, the synthetic data are assumed free of measurement errors throughout this work". I can certainly see the advantage of doing this in order to separate out analytical errors from other errors, as discussed in the paper. The authors mention in the end of the paper that analytical uncertainties (and a few other issues) will have to be considered as well. However, the authors spend a considerable amount of text on discussing different eMLR models that best represent the "true" change in DIC. I feel that without any quantitative discussions on the role of analytical uncertainties, the value of this discussion is reduced. Obviously, analytical uncertainties varies between cruises, but for an analysis where, for instance, GLODAP data is compared to CLIVAR data the average uncertainties of the various variables is relatively well known so that this could be factored in into the analysis.

As the reviewer points out, a complete analysis of uncertainties will have to be performed at one point or another but this task, which is not trivial, is beyond the scope of this paper. Measurement uncertainties are only one of many sources of error that will have to be addressed. The propagation of uncertainties is also related to the problem of independence in inverse problems (non-orthogonality of the vectors in the design matrix leads to the issue of matrix conditioning). Section 2 (eMLR theory) provides some elements of answer by introducing covariance matrices to the eMLR literature, but populating these matrices, particularly the off-diagonal elements, and evaluating how these may change temporally, will require much care.

Now, the reviewer has a valid point and it is true that measurement uncertainties (irrespective of other sources of errors) have an effect on model selection in regression analysis and it is possible that consideration of measurement errors affects the gradients in dynamic range (spread of the data normalized to the analytical uncertainty) between the tracers in the design matrices relative to these presented here. This should be considered when analyzing real observations but was not included in this model-based study for multiple reasons.

First, other model-based eMLR analyses (and many MLR/eMLR analyses, in fact) do not consider this effect either and we aimed to mimic previous eMLR implementations so as to isolate the influence of model selection and sampling network on eMLR results in comparison with these published studies.

Secondly, although we discuss the model selected and explain why some tracers may be selected preferentially over other, we more importantly present ways in which different models influence final results and suggest criteria that can aid in model selection. These conclusions/recommendations remain valid even without considering measurement uncertainties. It should be kept in mind that we only worked with first-order linear models (again to mimic previously published MLR/eMLR analyses) in a model framework. Although we find reasonable explanation for our selected models and these agree with previous analyses, the readers should not feel that the particular models we end up selecting should not be replaced by better ones (possibly non-linear ones): the concept of "best model" is quite plastic and depends, amongst other things, on the data set used and sampling density.

On page 14610, line 16 it is stated that "Models with the lowest overall AIC values tend to be the more complex ones". Even though this is true, it would be interesting to know if this is also true when analytical uncertainties are weighted into the analysis.

Consideration of the measurement uncertainties is expected to decrease optimal model complexity since the value of L in the first term defining AIC (-2ln(L)) will likely be smaller and in consequence the effect of the penalty term will be relatively more important. Given the magnitude of the numbers in the design matrices used in this study and rough estimates of the measurement uncertainties, one would expect L to decrease by one to two orders of magnitude when normalizing design matrices to the measurement uncertainties. Although we have not considered the covariance matrices explicitly in our study (for reasons stated above) and so cannot address this point definitely, we have performed perturbation analyses of the penalty term in the definition of AIC. These perturbation analyses showed that the penalty term had to be increased by 2 orders of magnitude to steer selected models towards the 4-term category. Given that the first term considers the natural log of L, this is a large increase in penalty. Consideration of measurement uncertainties tends to bring the AIC results in closer agreement with results obtained solely by the continuity criterion (a good thing). This issue does not affect the main conclusions of our paper.

The sentence was changed to "In the absence of analytical uncertainties, models with the lowest overall AIC… "

**Minor comments:**

Page 14601, line 14: "Absolute errors due to mapping are small and result in maximum vertically integrated column inventory biases of order 10 mol $m^{-2}$". Agreed that 10 mol $m^{-2}$ is a small number in relation to the column inventory of DIC. However, this is about the same amount that can be expected as the decadal uptake of DIC in the north Atlantic, so it is really a relatively large number, i.e. an error of the same magnitude as the signal you are trying to determine. This might be significant since the station distribution is different between the "GLODAP" and "CLIVAR" stations.

The vertically integrated mapping error can be of comparable magnitude to the signal in some restricted but dynamic regions where there is no data (comparison of Figure 1c and d with 1a). Compared to the scale of the basin, however, these regions are small (for example the Mann Eddy, in the central Labrador Sea, Caribbean Sea) and contribute little to the basin-scale inventory. Throughout the basin, most of the vertically integrated mapping errors are quite small, between 0 and 5 mol/$m^2$, which is smaller than the signal. When integrated over the whole volume, as is shown in Figure 5, the mapping error is very close to zero. Comparison of Figure 1a and 1b also show that the main spatial pattern of the signal is not greatly affected by the mapping process. The absolute errors that stem from the eMLR process and shown in Figure 7 are much larger.

The text was changed to: "In this study mapping errors are small in relation to other sources of uncertainties and are only significant in restricted regions, where the vertically integrated column inventory biases can be +/-10~mol/$m^2$, about half the size of the anthropogenic signal (Figure 1c, d)."

Page 14609, line 8: Similar to the previous point; the integrated mapping errors might be small, but on a regional scale it seems to me that they can be significant.

Vertically integrated mapping errors can be significant locally, as shown in Figure 1c

and d. The focus of this work, however, is on the basin-scale inventory change. On these larger scales, mapping errors are small relative to other sources of uncertainties. Oceanographic data are sparse and more data is always desirable but in relation to other problems of the eMLR method, mapping is not a first-order concern. One should also point out that mapping error could be significantly reduced with carefully located stations. In that sense, Figure 1c and d can be used to help search for additional data or guide future sampling.

Figure 1d: It is difficult to understand and believe that the gridding routine for the CLIVAR station network can yield such a good fit with the "true" data for some areas. This is particularly the case for the Labrador Sea where the CLIVAR station net don't have any stations, yet the difference between the mapped and true DeltaC is relatively modest.

The Labrador Sea is a region with a high mapping error in Figure 1d, consistent with expectation since there are no data in this region. One should point out, however, that gridding routines do not provide a "fit" to regions with no data, but generate a prediction of what the most likely values in these regions are. Most gridding routines, such as the one used here, perform the interpolation using a sum of multiple terms. The first term is typically the mean of the data on the surface to be mapped. That term in itself is usually a good first guess. Additional terms in the sum are corrections to the mean term that take into account spatial gradients of neighboring points (whose influence is dictated by the scales of the spatial covariance used) to provide an estimate of the local trends in the field. These secondary terms tend to zero in regions far from existing data such that in these regions, the prediction is dominated by the mean term. If the "true" values in regions with missing data are close to the mean, the prediction will be quite close to the "true" values.

We have not plotted the uncertainties of the maps, but it should be kept in mind that in these regions with no data, even if the central krigging estimate is close to the true value, the error of the krigging estimator is large. In the cases we tested, the true value was generally within the error bounds provided by the krigging uncertainties.

## Response to reviewer #2

page 14591, line 4. Please replace ”Independant” to ”Independent”
Done

page 14591, line 19. Please change ”cylcing” with ”cycling”
Done

page 14591, line 26. ”Difficutlies”, spelled wrong.
Done

page 14592, line 11. Actually the eMLR type of implementation of Wallace's (1995) MLR method, wasn't first proposed by Friis et al. (2005) but by Sonnerup, R. E., P. D. Quay, and A. P. McNichol (2000), The Indian Ocean 13C Suess effect, Global Biogeochem. Cycles, 14, 903– 916. It would be appropriate to also cite this paper.
Thank you for pointing this out. The reference was inserted.

page 14592, last section states (indirectly) that the model represents adequately ”natural temporal and spatial variability patterns that are consistent with many observed climate processes on a variety of time and space scales.” This is actually quite critical as the many studies from the NA, (as cited in the paper and also Metzl et al., BGC 2010) have shown that natural processes introduces significant carbon cycle variations over a few years. This statement must therefore be substantiated, and it would be a natural component of this paper to show that the model reproduces at least some of this and that the eMLR approach are able to distinguish between these quite significant variations and the anthropogenic imprint.
The model is forced with CORE fluxes, an observationally-based product of surface fluxes. The CORE product thus includes the dominant patterns of observed variability. The ocean model responds to these perturbations but does not feed back to the atmosphere. As such, the ocean model includes many natural perturbations in its physics that are forced by the observed atmospheric variability, to the limit that the ocean model construction and spin-up allow. The biogeochemical model (and so all tracers used in eMLR) will respond to these physical perturbations as prescribed by the model equations. There is no climate-carbon feedback in this model although this would be a small term on the time-scale considered.

In this study, the model is primarily used as a data generator to provide a data set that includes natural variability with a magnitude and patterns that are close to observations. Model performance is a secondary concern. A study investigating how different variability patterns affect the ability of eMLR to recover the true signal is an interesting proposition but it is beyond the scope of this paper. Here we have limited the discussion of model performance to section 4. This model setup has also been used and discussed in previously published works (for example Henson et al. 2009). These papers, listed later in the main text, also contain more information on various aspects of the model's performance.

The sentence was modified to “The model is forced by observed surface fluxes and so provides a means of estimating absolute errors in the presence of natural temporal and spatial variability patterns that are consistent with many observed climate processes on a variety of time and space scales.”

page 14594, following eq (2), define also “T”
Done

page 14597, line 3-4. The synthetic data are assumed free of measurement errors. As the authors should certainly be aware, this is not quite true. All measured data carry error, and some types of data carry more than others. A first-order analysis of

the effect of this issue on the selection of MLR is certainly in place, in order to unravel whether the effects discussed in the paper are inferior to, similar to, or larger than the effects of measurement error and how these vary from variable to variable.

The study of uncertainties is a very serious task and measurement errors are only one aspect of the uncertainties that should be considered. Measurement errors should be weighted against other sources of errors such as short-term variability of the tracers, non-orthogonality of the vectors in the design matrix and against the assumption inherent in this work that nominal years (i.e. 1995 and 2005 here) are representative of the data set. The results presented here required the calculations of tends of thousands of maps, a computationally intensive process. A careful account of the uncertainties will likely involve a Monte Carlo approach which implies repeating at least part of the analysis presented here many times. We strongly agree that a full consideration of the uncertainties will have to be done, but this will require a lot more work for it to be done correctly. We regard this as a necessary step in the implementation of eMLR to real observations and anticipate addressing this problem in future studies.

page 14598, line 16. Please provide some details on the "prescribed atmospheric carbon dioxide transient boundary condition". If I get this right, the atmospheric CO2 concentrations since the preindustrial are used as boundary, and not the emissions. This will introduce too little anthropogenic CO2 into the earth system as terrestrial and ocean uptake are already (implicitly) accounted for. Please discuss how this may affect the results that are presented.

Yes, atmospheric concentrations are prescribed, not emissions. As stated earlier, however, the primary use of the model here is as a data generator that includes temporal and spatial variability patterns reminiscent of observations. Model performance is not the focus of this work. Nonetheless, the model is state-of-the art and represents many features of the anthropogenic carbon inventory when compared with observational estimates as Figure R1 (below) shows.

We have only performed this analysis in one model and in one period (1995-2005). Repeating this analysis with other modeling systems and with other boundary conditions may be interesting, as model behavior will affect the magnitude of the variability in the model. Given the sizes of the errors in the eMLR calculations, we do not expect that differences between emissions and concentrations boundary conditions to be a first order problem nor that this will change our conclusions regarding model selection and the role of changing the observational network, however.

The following sentence was added: "The model global anthropogenic carbon inventory in 1995 is 104.9 PgC (Plancherel, 2012), a value within errors of the observational estimate (Sabine, 2004)."

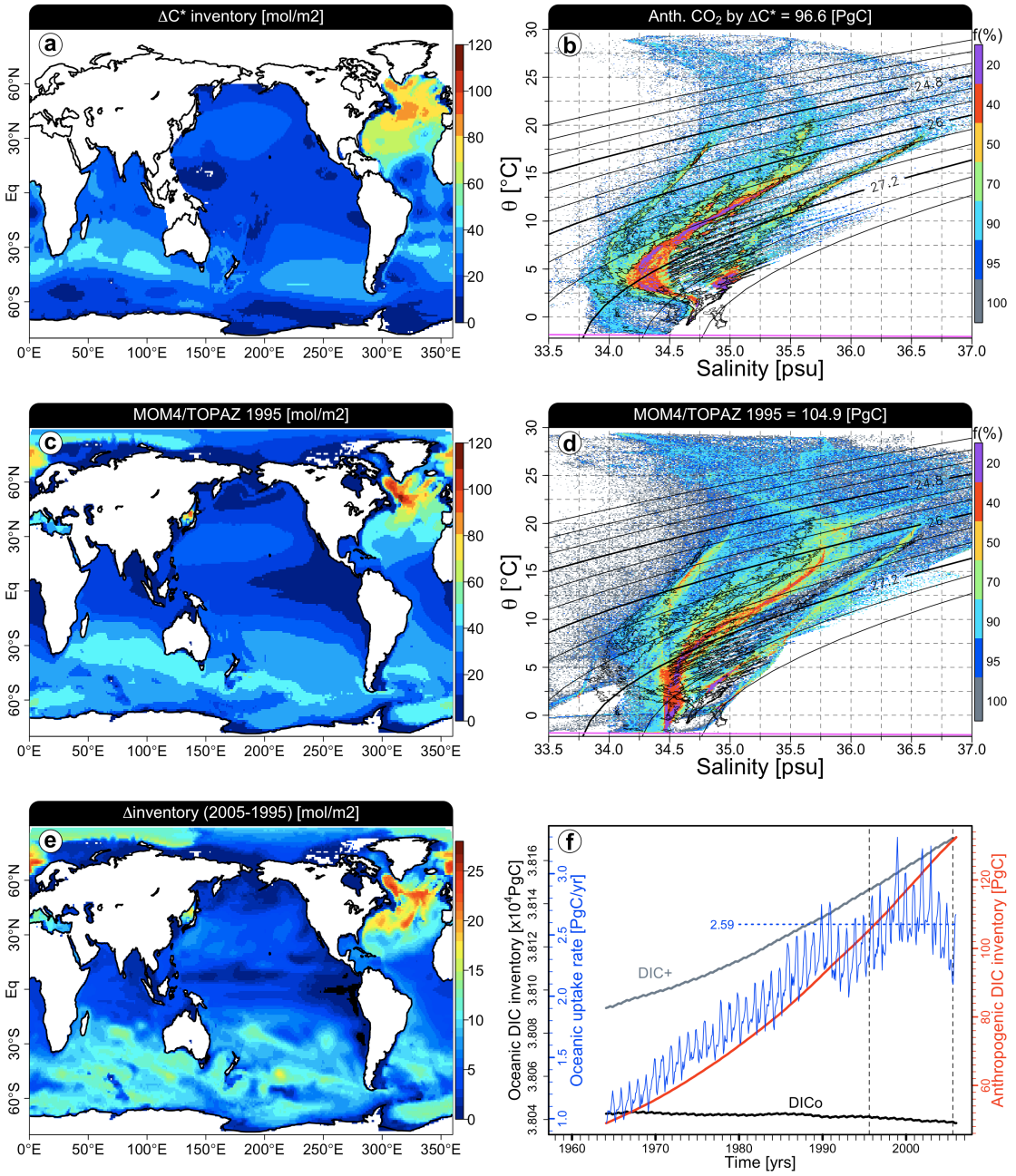


**Figure R1**: a) Anthropogenic carbon column inventory computed from the objectively mapped ΔC* GLODAP fields (Key et al., 2004). b) "Massic" θ/S diagram showing the 3D GLODAP ΔC* inventory projected onto the WOA05 annual climatology of potential temperature and salinity. The color bar indicates the relative cumulative anthropogenic carbon inventory sorted from the θ/S pair with largest to smallest inventory. Thin black contours show the outline of the corresponding volumetric θ/S distribution. c) Anthropogenic carbon inventory for July 1995 as simulated by the CORE-forced MOM4/TOPAZ model. d) Same as b) but with the 3D MOM4/TOPAZ inventory from c) and projected onto the model's own potential temperature and salinity fields. The thin black contours reproduce the outline of the volumetric θ/S distribution of WOA05 from b). e) Anthropogenic carbon column inventory difference between July 2005 and July 1995. Negative values are shaded in black . f) Evolution of the global DIC inventory in the control (DICo, black) and transient (DIC+, grey) simulations and associated anthropogenic carbon inventory (red) and uptake rate (blue). Vertical dashed bars are drawn in July 1995 and 2005. The mean uptake rate between 1995 and 2005 is indicated (2.59 PgC/yr). [Figure reproduced from Plancherel (2012). A study of the ocean's water masses using data and models. PhD Dissertation, Princeton University.]

page 14599. Regarding the Akaike Information Criterion. I am not all that familiar with this term, nor do I expect the audience of this paper to be. Please provide some information of what is considered good and acceptable values and what is not, so as the reader may get a bit more intuitive understanding of the implications of the numbers that appear in the paper.

As defined on p14599, AIC=-2ln(L)+2k. From this definition, AIC values are controlled by both the misfit to the data (the first term) and a penalty. Without the penalty, the value is simply a measure of the misfit. That is, ignoring the penalty term for now, AIC is large when the fit is poor. Better fits can always be achieved, however, by increasing the number of terms in the regression model, to the limit of the number of terms approaching the number of data points. This problem is commonly known as the variance-bias trade-off. In that sense, misfit by itself is not a sufficient measure of regression quality since low misfit (perfect fit, low variance) can always be achieved by increasing model complexity (increasing the model bias). The penalty term, here "2k" finds its rationale from information theory. Its function is to add "artificial misfit" to the direct measure of misfit (the first term) to penalize more complex models in a rational way.

Given a sequence of model of increasing complexity, a plot of AIC (on the y-axis) as a function of the number of model parameters (on the x-axis) would show a parabola. The models around the minimum of the parabola are those that fit the data well without resourcing to excessive complexity and thus provide a good consensus between fit quality and model complexity.

The text on p14599 was modified to make this more explicit and the following sentence was added: "For a given set of data, models corresponding to the smallest AIC values represent the best consensus between fit quality and model complexity."

page 14603, line 25-26. "Continuity of the regression formulae should thus be used as a model selection criterion in addition to standard measures of statistical misfit." Would it be possible to express this "continuity criterion" more explicitly?

Unfortunately, we do not yet have a "recipe" to quantify continuity and it remains a subjective measure, mostly based on the analysis of Figures 2 and 3. In our analysis, we used arguments about the distribution of water masses and we investigated the evolution of the model terms to rationalize it. Further research is necessary to better constrain and quantify this. At this point, we prefer to leave it as a concept.

page 14605, line 24. "Overall, nitrate and AOU were the least often selected variables in the formulae." Please discuss, briefly, the reasons for this preference.

With the caveat that measurement uncertainties are not considered in this analysis, the simple answer is that other variables are typically better explanatory variables on given depth layers. We do not have definitive physical arguments for this behavior but one can speculate, however, that nitrate, which is affected by anaerobic biogeochemical processes and nitrogen fixation may not be as clearly associated with DIC on the basin-scale as phosphate. Regarding AOU, its calculation involves an assumption of saturation. Preformed DIC has a degree of disequilibrium which is bigger than that of oxygen, which may explain why $O_2$ (without the saturation assumption) may be a preferred explanatory variable in some cases.

The following text was added: "This may be due to the fact that the distribution of nitrate is affected by denitrification and nitrogen fixation resulting in phosphate being the preferred variable. Similarly, there is an assumption of saturation in the

calculation of AOU, which may be incompatible with the DIC distribution in linear models. "

Figure A1 in the paper, showing the standard deviation of the data on each layer, also provides some element of answer. There are only few layers in Figure A1 where the standard deviation of AOU or nitrate, in relative terms, is greater than other tracers. Figure R2 below, taken from Plancherel (2012) summarizes the frequency with which explanatory variables are selected in the models as a function of depth and sampling network. We emphasize that these variables were "least" often selected, not "never" selected. As shown in Figure R2a and c, the difference between these tracers and others is not very large and is more dramatic for CLIVAR.

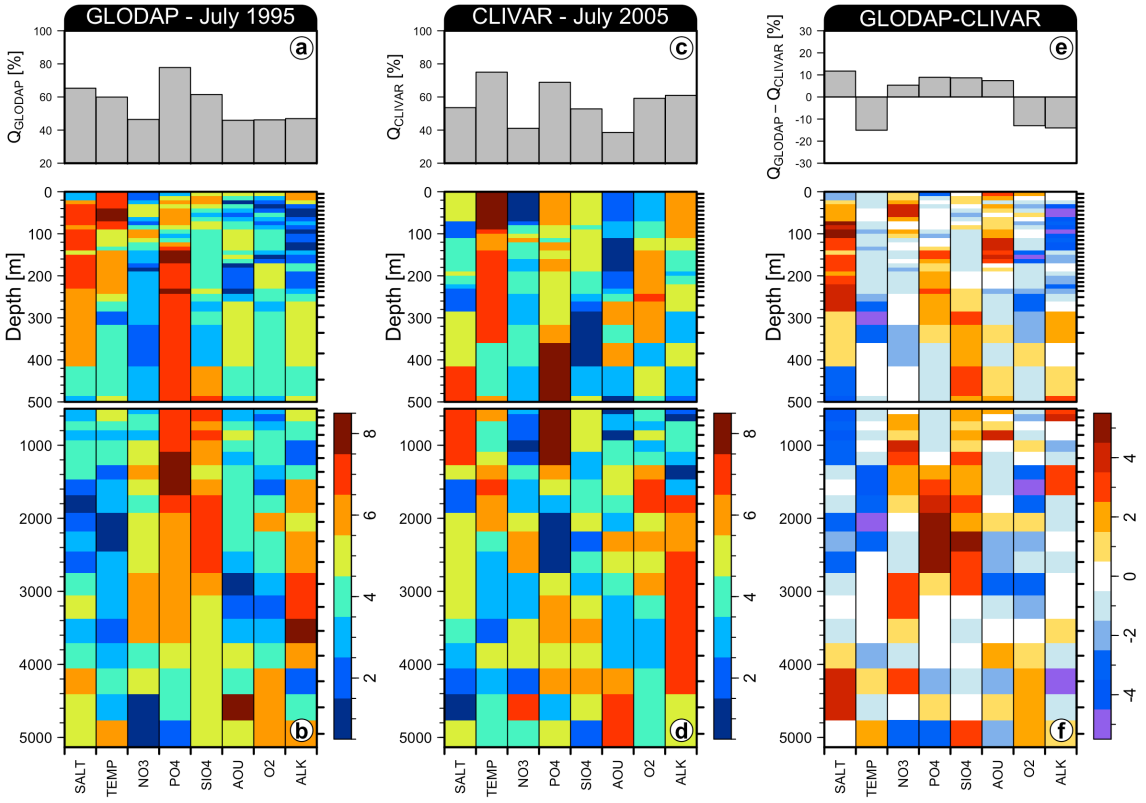


**Figure R2**: Summary of the frequency of occurrence of the variables from the starting set Z in the formula of all the models yielding minimum AIC values given either the 1995 GLODAP (a, b) or the 2005 CLIVAR (c, d) sampling grid and the difference between the two (e, f). b, d) For each depth layer, the color scale indicates the number of times that a variable is present in the formula corresponding to the model with minimum AIC in each size class. a, c) Total occurrence of the variables, summed over each layer, normalized to the maximum possible occurrence: $Q_j = \frac{\sum_{i=0}^{D} v_i}{8D}$, where D is the number of vertical layers. [Figure reproduced from Plancherel (2012). A study of the ocean's water masses using data and models. PhD Dissertation, Princeton University.]

Although not pointed out by the reviewer, a reference to Plancherel (2012) was also added on p14604, l13 instead of the previous "not shown" statement and that sentence was reworded to be clearer: "Parallel analyses for the complementary July 2005 GLODAP and July 1995 CLIVAR cases indicate that time exerts a smaller influence on the selection of regression formulae than changing the observational network (Plancherel, 2012)." We also removed "Futhermore" in the next sentence.

We also added "In these analyses" at the sentence beginning at p14606, l20 to better restrict and qualify the term "interannual variability" later in the sentence and its relation to the data set considered. "Interannual variability" should here be interpreted as 1995-2005 and the data set are GLODAP and CLIVAR. The statement in that sentence may differ for other times scales and other data sets.

page 14608, discussion of effects of seasonal cycle. The seasonal cycle of the errors is not fully aligned with seasonal MLD and PP cycles. Initialization of spring bloom occurs typically in April, May, June, and from the arguments presented in the paper I would expect greater variations/errors in these months. Please explore seasonal cycle of model in order to ensure that it is well aligned with the seasonal cycle as retrieved from various data sources.

Figure 5 shows that the errors are in fact large in April, May, June. Seasonality in this model, particularly the bloom cycle in the North Atlantic, was already investigated in Henson et al. (2009), which is listed as a reference in the paper. Henson et al. (2009) also show that there is no bloom in the sub-polar region for the period of September-October-November, the period of lowest error in Figure 5. This is also a period when the spatial standard deviation is high in most tracers (as shown in Figure A1).

In the text "late winter and spring" referred to all the processes listed. This was confusing as written and these words were removed from the sentence.

Figure R3, below, shows graphically the seasonal evolution of the regression misfit achieved in the best-AIC case. This shows that linear models have a much tougher time fitting the basin-scale data (near the surface) in summer than in winter. The reasons for this are not totally clear.

Although bloom dynamics plays a role, as it is a dominant cause of basin-scale seasonal variability, a direct link with the seasonal cycle of the error in Figure 5 is not clear. Physical and biogeochemical dynamics affect spatial tracer gradients, what in turns affect regression misfit. But since the domain is not homogenously sampled by the oceanographic stations, it also matters which station are affected most and how many of these stations are present in the data set that is being fitted. The seasonal cycle in Figure 5 thus represents the convolution of various effects.

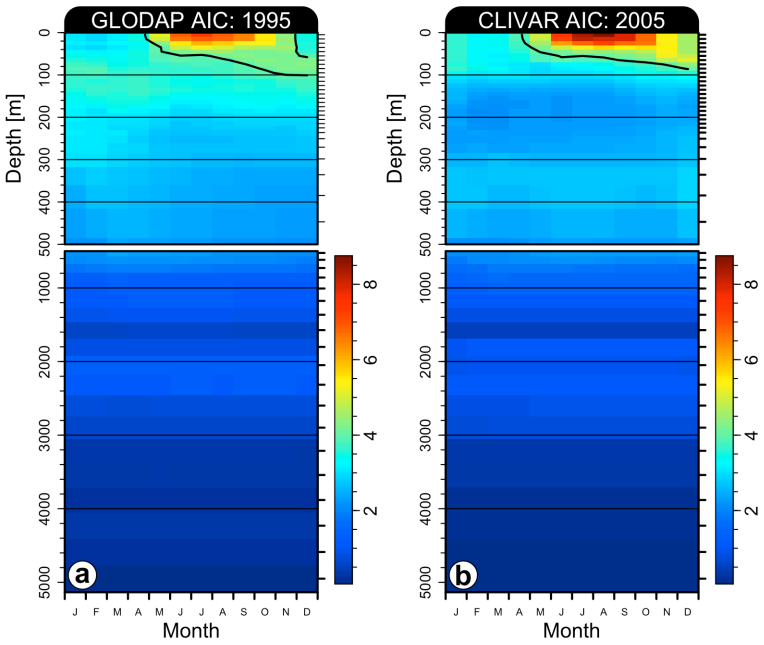


**Figure R3**: Seasonal evolution of the standard error of the residuals in μmol/kg calculated after application to each depth layer of the corresponding minimum AIC models isolated from all first order formulas for a) GLODAP 1995 and b) CLIVAR 2005. A black contour line in a) and b) is drawn at 4 μmol/kg, which is an estimate of the modern measurement accuracy of DIC. [Figure reproduced from Plancherel (2012). A study of the ocean's water masses using data and models. PhD Dissertation, Princeton University.]

page 14612. lines 25-30. Actually, it appears from the figure that the AIC produces the best accumulated profile.

Yes, but this does not exclude the fact that a composite solution can provide an even better solution. Moreover, the good vertically integrated profile is only "good" near the surface, at the end of the integration and this is because of cancellation of errors in the interior. For example, below 1000m, Z140 is far superior to the best AIC case.

page 14613, lines 2-4. Might a more rigid procedure for selection be suggested?

At this stage, we are unfortunately not able to provide a more objective/quantitative "recipe". Continuity is a concept at this point and the researcher's best judgment together with oceanographic knowledge is a necessary component of the procedure.

page 14614, line 21, please replace "formlae" with "formulae"

Done

page 14616, lines 8-10: "Given the available sample distribution, differences in representativeness of the sampling grids have a larger effect than changes in the seasonal distribution of the samples." This statement should be substantiated, by referring to a figure that actually shows this. From Fig 5, it appears as if the opposite is the case.

The difference between the GLODAP (magenta) and the CLIVAR (green) lines in Figure 5 only represents differences due to mapping the best-AIC eMLR results from these sets of stations. The offset between these two lines does not represent the influence of changing the observational network on model selection, which is the point that we were trying to make here. In this sentence we are trying to estimate if having a data set with points from various months would influence the regression model selection process more than the difference in sampling network.

The text was changed slightly to clarify this and a reference to Figure A2 was added.

page 14617, lines 11-13. I do not completely understand the argument made here. Please consider to make this more explicit.
According to the LSCOP criterion and the knowledge of the exact inventory change in the model, "acceptable" inventory change estimates should be between 0.343 to 0.543 PgC/yr (p14617, l11). There is no constraint on the source of this "allowed" error, however. It is generally accepted that this will come from the eMLR methodology only. Here we point out that rates derived from the eMLR result as performed here make another serious assumption, that is that the temporal distribution of the data set can be reduced to two nominal years, 1995 and 2005. What if the methodology was perfect (i.e. eMLR could reproduce the exact inventory change) but the problem came only from the assumption that exactly 10 years separates the data? As it is, it is not clear how to deal with the temporal distribution of the data on the basin-scale. This is in fact a serious shortcoming of the current eMLR approach.

Although the interval of 10 years was here based on the modal peaks present in the distribution of the data, our analyses have shown that data density is not the only factor; it also matters where each data is taken as this governs the influence that each data point has on the regressions. Let's imagine that a methodology can be developed that allows one to estimate the weighted influence of each point on the temporal aspect of the data set and that this gives an effective mean time difference between the data sets of 8 years instead of 10 years. Assume one gets an error-free eMLR estimate of the change in inventory given the data sets and that number is 4.43 PgC, for the sake of argument (the same as the perfect model inventory change). Given the magnitude of that number, if one were to calculate a rate by normalizing with the time difference, the rate will vary depending on the assumed time-difference between the data sets. If the time difference is exactly 10 years, the rate is 0.443 PgC/yr. If the "correct" time difference were instead 8 years, the rate would be 4.43 PgC/8 years = 0.55 PgC/yr. If it was 12 years, the rate would be 4.43 PgC/12 years = 0.37 PgC/yr (p14617, l9). These numbers are close to the boundary of "acceptable" errors suggested by LSCOP, i.e. 0.343 and 0.543 PgC/yr for this example. That is, a 0.1 PgC/yr error in the North Atlantic is about the same as a 2 year error on the timing assumption.

The point of the sentence on p 14617, l11-13 is to put these numbers in perspective and simply to point out that since our eMLR results tend to underestimate the true inventory change, if one were to divide this smaller number by a smaller time interval (instead of the correct 10 years interval), this could yield the correct rate. Because the methodology is not perfect, this biases the influence of the timing error. If the "true" characteristic time interval between the sampling campaigns (including somehow the spatial influence of the stations) were in fact smaller than 10 years, this would (by chance) correct the rate values. If, however, the characteristic time interval were larger than 10 years, because the eMLR results are underestimated, one would divide a small number by a larger number and the rate would be underestimated even more.

This section was reorganized and the text was changed to make this clearer.

Finally. The authors may consider adding a third strategy to their approach. Use of regression formulae with sets of explanatory variables that are allowed to vary as a function of depth, but not in time.

This is in fact what we propose in our “consensus” strategy and one of the points of Figure 8. In that case, vertical continuity and data set “representativeness” are used as arguments to aid model selection and fix the model structure in time but allow them to vary in depth.

## Response to reviewer #3

**Specific comment**

I would like to stress on three issues: the premises of eMLR described in the manuscript, seasonal variability and the vertical coordinates.

1.- eMLR assumptions. As the authors stated: "Assuming that suitable empirical regression models can be found for DIC and that the physical and biogeochemical processes underlying the model are stationary and not affected by the anthropogenic signal, the noise can be filtered out and the anthropogenic signal revealed as the difference between model predictions of DIC at different times (Friis et al., 2005)".

However authors show several features that contradicts this assumption. Several figures show that Cant and the anomalies of Cant show some correlation with the natural CO2. For instance, in Pag 14610 lines (21-26) authors claimed : 'The error pattern of Fig. 7g–i look similar to the column inventory change pattern resulting from the natural carbon run (Fig. 1f), although with absolute errors of overall smaller magnitudes than the vertically integrated natural carbon change. This pattern similarity indicates that, while eMLR accounts for some of the natural variability, large scale natural variability patterns are not fully corrected for, even when the statistically best models are used systematically".

This fact is associated mainly to the natural variability of LSW in the layer between 1500 and 3500 m, where most of the errors (underestimation, Fig. 6) in the basinscale inventory change estimates are found. The authors find [Page 14612 (lines 14-21)] a notable correspondence in a layer with high values of AIC around 2000m which corresponds to a layer of systematic underestimations of Cant (Fig. 8a). The underestimation in this layer generates most of the error in the basin-scale inventory change estimates. In this layer, the negative biases of the Cant change inventory (yellow lines vs black lines in Fig8) are associated with the negative layer inventory change of the natural carbon simulation (green line in Fig. 8). The authors claim for "…clearly shows the effect of the Labrador Sea Water variability and water mass reorganization…", recognizing that the physical processes are not well isolated in the eMLR model affecting to the Cant change estimates.

This issue is one of the main disadvantages of applying eMLR in the North Atlantic. Kieke et al (2007) showed the strong variability in the CFCs content in the Labrador Sea Water related to the NAO variability. Steinfeld et al. (2009) evaluated the changes in the Cant using CFC data and the thickness of the main water masses in the North Atlantic and showed a clear decrease in the Cant storage rate due to the volume reduction in the LSW classes between 1997 and 2003. Also, Perez et al. (2008) showed a decrease of the Cant storage rate related with the NAO variability and described a strong correlation between Cant and O2 in the main water mass (Johnson et al. 2005).

This relationship is contradictory to the assumptions needed to apply the eMLR technics. But this inconvenient is not only observed in relation to LSW, also Rodgers et al. (2009) showed that in the North Atlantic, the top 1000m vertical integration of O2 and total CO2 does not show any spatial co-variability which suggests that natural and anthropogenic CO2 concentration should have a negative correlation. All this evidences and those described in the manuscript indicate that the observed decadal variability of the physical drivers in the North Atlantic between 1995 and 2005, associated with the NAO, are driving changes in both the natural component of the carbon cycle and also the anthropogenic component, generating a cross correlation in both component in contradiction with the assumptions of the eMLR technique.

Yes, natural variability is an important feature of the North Atlantic and of the ocean in general. The confounding effect of this variability on the distribution of tracers, and for this application, anthropogenic carbon, is the main motivation to use a regression-based method to begin with. The expectation is that processes governing the natural variability will continuously exist in time and so be captured by the earlier regression. This is an inherent assumption of the technique. The degree to which it is true was not clear and that is one of the motivations why we used a model that includes natural variability to test the eMLR method. The model does not include a carbon-climate feedback, but we expect these effects to be small on the time-scale considered. We also agree that some results show a spatial correlation between error patterns and the natural carbon changes. As stated in the text, these show that eMLR is not able to remove all the effect of natural variability. The degree to which eMLR is successful depends on which regression model is used. We are unsure what the reviewer's criticism is regarding the existence of natural variability,

but we hope that our new abstract and conclusion section addresses some of them.

As the results presented in the paper show, as is acknowledged in the paper and quoted above, we recognize that eMLR is not a perfect solution in that maps of the vertically integrated errors can show patterns similar to the natural variability. This also goes to show that volume-based inventories may not be very stringent criteria of quality as they are relatively easy to achieve. As described for instance in section 6.1.2, many of the results meet the LSCOP criterion, yet maps of errors show spatial patterns, emphasizing the point that a volume-based diagnostic is not a sufficient constraint.

2.- Seasonal variability Why estimated Cant does vary seasonally while the 'true' signal is constant? Authors argued that it originates from the seasonally varying ability of linear models to fit the data. However, a negative bias is found again in the estimated Cant signal, mainly in late winter that could be related to the spatial cross correlation variabilities in the O2, natural CO2 and Cant, which seem to be lower in summer and autumn considering the lower AIC numbers. In fact authors noted 'since regression misfits are largest in the summer and early fall, addition of winter and spring data should result in an overall improvement of the fit quality,' suggesting that better AIC should produce negative biases in the estimated Cant changes because part of the Cant signal is included in the nominal natural signal of the eMLR model.

The true signal also varies seasonally (see Figure 5 and R1 above). The true seasonal cycle is smaller than that the eMLR results in Figure 5, however. That later cycle is modulated by errors in the application of the eMLR method. Our analysis of maps of the tracer fields in the model and investigation of the evolution of the variance on depth layer indicates that these latter considerations dominate the seasonal differences of fit quality.

Correlation in variability of the tracers is in fact a desired feature at the root of eMLR. If these correlations are constant in time, the assumption is that the empirical regression model can predict DIC accurately given the other tracers have varied. The problem arises when these correlations fall down.

The problem of cross-correlation between tracers is mentioned in our discussion (we use the terms "optimizing tracer orthogonality"). This is a serious issue that will require some attention in the future. How this problem affects the eMLR results has never been quantitatively assessed. The most directly relevant paper on the topic of tracer orthogonality we know is that of Fukumori and Wunsch (1991) Efficient representation of the North Atlantic hydrographic and chemical distributions. Prog. Oceanog. 27, 111-195.

3.- Vertical coordinates. Authors performed this statistical analysis in horizontal surfaces because of the design of the model outputs. However, they have suggested that the results would be better if the analysis were performed in isoneutral surfaces. It is certainly true that the analysis would be more precise because the variability along isopycnal is dramatically lower than use horizontal surfaces along the whole North Atlantic.

In addition, most of the recent works using eMLR did perform the computations over isopycnal layers, so it seems rather evident that the objectives of the manuscript would be better addressed if the analysis were done over isopycnal surface rather than horizontal surfaces from the beginning.

Yes, an application on surfaces that better mimic the dynamical constraints of ocean mixing and circulation will help and we acknowledge this in our discussion section. Since gradients will be smaller and mixing better represented on isopycnals, eMLR is expected to perform better on these surfaces, with errors smaller than these presented here. The most recent calculations with observations make use of

isopycnals but all previous model-based assessments of the method are performed on depth surface. Overall, even if an isopycnal-based analysis will likely affect the models selected and the quality of the solution (for the better), the main points, that is the use of continuity as a criterion selection, the fact that changing the observational network has a strong influence on the selection process, and that a basin-scale inventory change can be achieved with good accuracy will remain valid.

A basic analysis of how mixing and biology affects the eMLR results is available in section 5.5.2.1 in [Plancherel (2012). A study of the ocean's water masses using data and models. PhD Dissertation, Princeton University]. This discussion was not included in the present paper as this is not directly relevant to the results presented here.

**Other specific comments**

Authors also suggest the use of using quasi-conservative variables as independent ones in order to optimize the tracer orthogonality. However, the lack of orthogonality is more relevant between natural and anthropogenic CO2 that produced a negative bias in the estimation of the anthropogenic signal. Following the idea of the authors about quasi-conservative variables, it would be more practical the use of C* or TrOCA parameters as dependent variables to determine the anthropogenic signal instead of using the observed contemporary total carbon to evaluate the decadal change. C* or TrOCA parameters are nearly independent of the natural carbon cycle, thus avoiding the cross correlation observed in the North Atlantic between Cant, natural CO2 and oxygen.

The use of star tracers is an interesting proposition and we have also suggested this in our discussion. Quite a few assumptions go into the derivation of these tracers, however. At this point, it is not clear if what is gained in terms of problem conditioning (and hence error propagation and solution stability) by using these star tracers or TrOCA is not balanced by the assumptions necessary to compute these tracers. A brief comparison of the concepts behind the eMLR method with the $\Delta C^*$ method is available in section 5.2.2 in Plancherel (2012) and the discussion about mixing and biology mentioned above (section 5.5.2.1) may be useful in the future when evaluating the use of star-tracers.

## Response to reviewer #4

1. While this paper provides an in-depth analysis of different regression models and the impact of models with different numbers of parameters, I feel that the discussion and conclusion are in need of some stronger statements as to the findings of these analyses. Specifically, how can field studies use these findings to determine optimal eMLR regression models? The authors allude to this on page 14613 and in the discussion but the reader comes away unsure how the findings from this study could be directly applied to field data where the 'true uptake' rate is unknown and so a rigorous evaluation of regression models is difficult.

Aside from pointing out that an accurate estimate of the basin-scale inventory change is possible by eMLR, our work demonstrates an analytical process including the problem of regression model selection (aided by oceanographic knowledge and the concept of continuity) and highlights the role of varying the observational network. In that sense, field studies can use the tools we presented (mostly the diagnostic tools, such as Figure 2 and 3) and the theoretical framework. The abstract and the conclusion section were rewritten (see below) and some aspects of the discussion were amended/expanded.

Our work should be regarded more as an example of an application than a definite recipe. Many issues have yet to be addressed regarding eMLR. Although we present a few regression formulae and these regression formulae agree with other analyses in the North Atlantic, it would be unwise to use these blindly when applying the analysis to data. There are many assumptions behind our model-based work and it is likely that an investigator using a data set different than the one used here will want to use different regressions. This will also be the case if the analysis is performed on isoneutrals, as we recommend doing. As pointed out by other reviewers, we have also not considered measurement uncertainties. These have the potential to alter the model selection.

(a) Is there a relationship between the MLR residuals (or AIC values) and the delta Inventory relative error? This analysis would be simple to add as the data already are presented in Figures 4 and 6.

This is an interesting proposition, but the required analysis is not so simple, in fact. MLR residuals exist for each data point and AIC values exist on each layer. On the other hand, the error on the change in inventory, as proposed, is a single number for the whole basin. Also, there are two kinds of residuals, one for each time period and sampling network (GLODAP and CLIVAR). Goodkin et al. (2011) talked about the cancellation effect of the residuals in a fixed sampling network and fixed regression framework, showing that this effect is in fact an important reason why eMLR seems to work as well as it does. Here it is difficult to make general statements about the cancelation effect as both the regression and the sampling network are allowed to change and influence the residual field.

We can comment on the spatial distribution of the residuals, however. An analysis of the spatial distribution of the residuals and of the influence of each station on the regression was performed in section 5.4.3 in Plancherel (2012) [A study of the ocean's water masses using data and models. PhD Dissertation, Princeton University]. Maps of the residuals show that the residuals are not uniformly

distributed geographically, even for the most complex models. There exist spatially coherent domain of positive and negative residuals that vary depending on the regression model used. This indicates that what is considered as "noise" in the regression process is not simply due to the eddy field, as one would expect eddy-induced variability to show more spotty patterns. Given a fixed regression model, or even closely related models, stations with extreme residuals tend to exist in the same regions. This indicates that a portion of the errors arising from the regression misfits tend to cancel in the eMLR calculation. More work is needed to understand how the spatial distribution of the residuals from variable sampling network influences the inventory calculation.

The text relevant to this discussion in Section "7.3 Spatial sampling density" was expanded to address this question. The text now reads:
"Because of inhomogenous and non-random sampling of the ocean, an eMLR implementation based purely on statistical arguments (i.e. ``best AIC'', strategy 1) will not necessarily yield the most accurate answer. This is because local features may be present differently in each regression fit as these are derived from different sampling networks. These differences in fit quality influence the information contained in the residual field and so affect the empirical definition of ``natural variability''. This can be interpreted as a kind of over-fitting, although not in a statistical sense specific to each regression individually but in a pragmatic sense, with respect to the eMLR process as a whole.

When using the same formula in time (strategy 2 and the composite strategy), structures due to regression misfit are more likely to cancel. A quantitative assessment of this effect is difficult when the sampling grid varies, however. Analysis of the geographical distribution of the residuals indicates that misfit is not homogenous but that residuals can form large spatially coherent patterns (Plancherel, 2012). These patterns tend to be analogous for similar regression formulae applied to the different sampling grids but the degree of cancellation varies with formulae and with region. The final influence of this cancellation effect on basin-scale estimates depends on the size of the regions after propagation through the mapping routine as well. "

(b) The authors find that regression models with 7 parameters are often selected. However, for many observational datasets this large number of parameters is not available. Could the authors expand on the analyses presented on page 14605 and 14606 to show (perhaps with an additional figure or table) which parameters are most commonly selected for and so should be a high priority for programs such as CLIVAR?

The tracers we considered are standard measures on CLIVAR cruises (i.e. T, S, nutrients and carbon system parameters) and are already considered high priority. In fact, the least available data are those pertaining to the carbon system parameters (specifically alkalinity in the real data set given the set of tracers considered here). Since DIC is a necessary variable for this application, however, it is largely this variable that is the limiting one. At the beginning of this work, we had actually also included other tracers, but we ended up limiting the analysis to the standard tracers presented here precisely because of concerns on data availability. Future work should certainly explore the use of other variables, however, especially as tracer

independence is an issue.

Similarly, the authors mention that some parameters often replace each other such that it would not be necessary to include both in a regression model. Could this information be expanded upon so that all of these related groups are presented to the reader instead of just a select few?

The limited discussion on p14605, amongst other considerations, comes from the analysis of Figure R2, above. Figure R2 shows the frequency, across all 255 models, with which a particular tracer is selected in the regression selection process. We have opted to leave this figure out to limit the overall length of the manuscript, however. The few groups we talked about are the dominant ones. Given the limitations of our work (it is model based, it ignores measurement errors and the assumption that the data sets can be adequately reduced to two nominal years) and the fact that the data set will likely evolve between our initial estimate of what constitute the CLIVAR stations and the final version that will be use for real calculations, we feel that extending and focusing the discussion on the exact structure of the regression formulae would not be very helpful. Ultimately, we feel that an expanded discussion of this term-swapping effect will be more in its place in a study investigating the issue of tracer correlations and problem conditioning.

2. The authors present and analyze two eMLR strategies in addition to a combined strategy but do not clearly state which strategy they find evidence for being the preferred strategy. Similar to my comment #1, this leaves the reader unsure how these findings could best be applied to other studies. The comprehensive analysis done by the authors should make a set of guidelines for 'strategy selection' easy to provide to the reader.

The conclusion section was rewritten entirely to highlights the main results more clearly. The new section is as follows:

"The eMLR method was evaluated using output from a global circulation and biogeochemistry model with a known anthropogenic signal and representative spatio-temporal patterns of variability from which absolute errors could be evaluated. The model was sampled at observed station locations to create synthetic data sets that mimic the spatial structure of the observed data. Recasting the eMLR equation in the formalism of inverse problems allows one to consider different application strategies with regression models that can change in time.

Analysis of these synthetic data sets shows that both the station distribution and the selection of regression models exert strong influences on the eMLR's ability to recover the true signal locally. Interior layer inventory change estimates can err by as much as 100% or more when the analysis is performed on horizontal surfaces and uses inappropriate regression models, even if the basin-scale inventory change is in agreement with the true value. The depth range between 1500 and 3500~m is particularly sensitive to model selection and is the source of most of the difference between the true signal and the eMLR-inferred signals in the North Atlantic.

The model selection process depends on the station distribution. Sampling density influences data set variance and fit quality, *de facto* modifying the type of natural variability captured in the residual field. The systematic use of the statistically optimum regression fits at all time and depth layers does not guarantee the best results if the sampling network emphasizes regions differently in time. Consideration of spatial representativeness of the sampling network and vertical

continuity of the regression formulae as justified by oceanographic knowledge are key in selecting the appropriate regression models. Best results are obtained when regression formula are allowed to vary in depth but not in time. Applying these guidelines, and given the station location considered in this model-based analysis, we find that stacking regression formulae $Z_{100}$={S,θ,$PO_4$,$O_2$} in the top 1000m and $Z_{140}$={θ,$PO_4$,Si,Alk} below that depth gives accurate results in the North Atlantic both in the interior and when integrated over the volume when the fits are performed on multiple horizontal layers.

The basin-scale application of eMLR presented here exhibits skill in detecting integrated ocean carbon decadal changes that fall within the threshold of acceptable uncertainty (10%) proposed in the LSCOP report (Bender et al. 2002). The fact that eMLR produces good results in the hydrographically complex and dynamic North Atlantic suggests that it will likely perform well in other hydrographically simpler and less variable basins. Although further development and assessment of the method is encouraged, particularly to address the issues of temporally variable covariances, problem conditioning and temporal staggering of the samples, eMLR presents an opportunity to exploit the many interior DIC measurements and evaluate the evolution of the ocean carbon sink and its rate of change independently from other estimates."

(a) Goodkin et al 2012 describe the theory behind the eMLR approach, specifically that by applying a regression model at both t1 and t2 the MLR residuals cancel each other out. It would be interesting to see whether this assumption holds for strategy 1 where two different regression models with different variables are used. One would assume that different regression models would produce different residual patterns. This could explain some of the differences in the $\Delta C_{anthro}$ estimate.

Given a fixed sampling network, and a fixed regression formula (as in Goodkin et al. 2011), the residuals vary slightly in time and so largely cancel. Given a fixed sampling network and a variable regression formula, the residuals vary more. Since the best-AIC models (strategy 1) are typically the most complex ones, however, the change in residuals between these fits is quite small. Residual patterns are much more variable between models that are less complex than for formulae that are more complex as these account for more of the variance. An analysis of maps of residuals indicate that the spatial patterns of the residuals vary most due to the variable nature of the sampling network, however. That is, given a constant field, a constant formula but a variable sampling network has the potential to yield relatively high differences in residuals.

We expect that our expanded section 7.3 (quoted above, response to comment #1a) addresses this question in the main text.

A full account of errors and their propagations has not been performed, but these considerations about residuals should surely be part of the discussion in any future study focusing on eMLR uncertainties.

(b) It is not clear why the results from strategy 1 are displayed in a different format (Figure 5) from the results from strategy 2 (Figure 6). It would be helpful to have directly comparable figures, if only in the supplemental material.

The seasonal variation shown in Figure 5 is quite small in relation to the spread in Figure 6. We opted to separate these results to emphasize the seasonal aspect in

Figure 5 and to not clutter even more Figure 6. The left y-axis in both figures shows the same quantity (relative errors), and it should be easy to imagine where the July results from Figure 5 plot in Figure 6 (at about -4% and -6%).

3. The authors do not comment on the possibility of model drift explaining some of the observed patterns especially in the control run (page 14602). The spin-up period for these model simulations was relatively short and so model drift, particularly in the deep ocean, would not be unexpected. This should have a relatively small impact on the primary analyses of this paper however it is an important limitation of the dataset that the authors used and should be mentioned.

An estimate of the global effect of model drift on the global carbon inventory in the control simulation can be see in Figure R1f, above. On the time-scale considered in this paper, model drift is not an issue and it is dwarfed by the influence of natural variability and especially by the effect of changing the observational network.

**Minor comments:**

Abstract: I felt the abstract does not do this paper justice. I would suggest shortening the sentences to make it easier to read, providing a motivating first sentence, and providing more specific conclusions.

The abstract was rewritten. The new abstract reads as follows:

"Quantifying oceanic anthropogenic carbon uptake by monitoring interior dissolved inorganic carbon (DIC) concentrations is complicated by the influence of natural variability on these concentrations. The ``eMLR method'' aims to address this issue by using regression fits of the data instead of the data themselves. The anthropogenic carbon fraction is calculated by evaluating the difference between predictions generated by the regressions at each time. The advantages of the method are that it provides a means to filter out natural variability, which theoretically becomes the regression residuals, and a way to deal with sparsely distributed data. The degree to which these advantages are realized in practice is unclear, however. Here, we investigate the ability of the eMLR method to recover the anthropogenic carbon signal using a global circulation and biogeochemistry model in which the true signal is known. We find that regression model selection strongly influences the magnitude and distribution of the recovered anthropogenic carbon change signal. The spatial distribution of the data exerts a dominant control on the regression results. The systematic use of statistically optimum regression formulae does not produce the best estimates of anthropogenic carbon change if the spatial distribution of the stations emphasizes hydrographic features differently in time. Important factors, such as a balanced and representative station distribution and vertical continuity of the regression formulae must be considered to guide model selection. Best results are obtained when multiple formulae are used throughout the water column. Although issues have yet to be addressed, we find that eMLR can estimate the North Atlantic basin-scale decadal change in anthropogenic carbon inventory with an accuracy better than 10%."

Pg 14592 line 11: "proposed to compare empirical" is not grammatically correct. Perhaps "proposed the comparison of"

Done

Pg 14592 line 18: "In preparation to an application of the eMLR approach to global data sets" needs to be reworded

Done, this phrase was removed as not really necessary.

Pg 14594 line 7: The following wording is difficult to understand: "predictions obtained from a model obtained from one data set"

The sentence was changed to "a set of DIC predictions generated from a model fitted to one data set but applied to the other data set is necessary".

Pg 14601 line 26: There is significant evidence from multiple studies that mode waters take up anthropogenic carbon in the subtropics (added to the list provided by the authors is an observational based analysis by Bates Biogeosciences, 2012 and a modeling analysis by Levine et al Global Biogeochemical Cycles, 2011). It seems that the lack of uptake by mode waters in this model represents a significant limitation that should be pointed out.

We do not doubt that the subtropics takes up anthropogenic carbon and the model actually shows this feature (Figure 1a). This statement referred to the transient run only (Figure 1e) but that was not very clear as written. We have reworded the paragraph as follows:

"Figures 1e and f show the vertically integrated changes in the transient and control simulations (the two components used to calculate the anthropogenic signal), interpolated from the set of samples taken at GLODAP locations. The July 2005-1995 difference in the transient simulation (Figure 1e, $\Delta C_{contemporary}$) reveals substantial carbon accumulation in the subpolar gyre region, the European Basin and at the southern edge of the subtropical gyre (15$^{o}$N) but apparently little change in the vertically integrated column carbon inventory in the region South of the Gulf Stream. Many of these features are compensated by the control simulation (Figure 1f, $\Delta C_{natural}$), highlighting the importance of natural variability. For example, the Western Subtropical Atlantic shows a drastic decrease between 1995 and 2005, which, when added to the transient run, results in substantial carbon uptake in the subtropical mode water formation region (Figure 1a, b), confirming expectations (Bates:1996, Lee:2003). The Greenland Current region, the Eastern Atlantic and the southern edge of the subtropical gyre all show increases in vertical carbon inventories in the control run (Figure 1f). "

Furthermore, it should be kept in mind that it is the vertically integrated column inventory that is being discussed. Subsequent paragraphs in this section explain that this feature is due to changes in the deep layers. The actual change in anthropogenic carbon is consistent with expectations in the upper ocean.

Pg 14603 line 15: "Observations" or "field data" might be more appropriate terms than "real data".

Done, changed to "field observations"

Pg 14621 line 5 "regressio" should be "regression"

Done

Pg 14622 line 6: "the variance is associated to a seasonal cycle" should be "the variance is associated with a seasonal cycle"

Done

Figure 2 & 3: I find these figures very difficult to read. I would recommend removing the background colors and displaying them as a colorbar on top of the figure as done in Figure 6.

We acknowledge these are complex figures. When removing the color background, it becomes difficult to see in which group the black bars belong to. When showing the black bars in their background colors instead, the yellow is hard to see. We find the background colors/black bars to be generally a good consensus.

Could the authors provide an appendix table which lists which parameters were used for each model number to allow for further interpretation of these figures?

We are happy to provide a list linking explicitly model number to model structure as supplementary material (Table S1). We slightly changed the text in the methodology section by adding "See Table S1 in supplementary material for a list of the model

terms."

Figure 4 caption: The magenta bars are almost impossible to see.

These are not easy to see, but they are reproductions from the white bars in Figure 2, where they can be more easily seen. Also, since the main point is the background colors and the black shaded areas, we feel this is a suitable, although not perfect, consensus. We've tried many iterations of Fig 4 with different color palettes (making the magenta bars white as was often suggested does not help). The version here is the best we could do.

Figure 6: Can the authors indicate the 'best fit' model for each model group?

There is not one best-fit model per group but a collection of them that vary with depth and sampling network. Figures 2 and 3 summarize this information graphically in various forms (in panels a, b and c).