Biogeosciences, 10, 4801–4831, 2013
www.biogeosciences.net/10/4801/2013/
doi:10.5194/bg-10-4801-2013

Biogeosciences Open Access

Role of regression model selection and station distribution on the estimation of oceanic anthropogenic carbon change by eMLR

Y. Plancherel[1,*], K. B. Rodgers[2], R. M. Key[2], A. R. Jacobson[3], and J. L. Sarmiento[2]

[1]Department of Geosciences, Princeton University, Princeton, New Jersey, USA
[2]AOS Program, Princeton University, Princeton, New Jersey, USA
[3]NOAA/ESRL/GMD, Boulder, Colorado, USA
[*]now at: Department of Earth Sciences and Oxford Martin School, University of Oxford, Oxford, OX1 3AN, UK

Correspondence to: Y. Plancherel (yvesp@earth.ox.ac.uk)

Received: 2 September 2012 – Published in Biogeosciences Discuss.: 19 October 2012
Revised: 3 June 2013 – Accepted: 17 June 2013 – Published: 16 July 2013

Abstract. Quantifying oceanic anthropogenic carbon uptake by monitoring interior dissolved inorganic carbon (DIC) concentrations is complicated by the influence of natural variability. The "eMLR method" aims to address this issue by using empirical regression fits of the data instead of the data themselves, inferring the change in anthropogenic carbon in time by difference between predictions generated by the regressions at each time. The advantages of the method are that it provides in principle a means to filter out natural variability, which theoretically becomes the regression residuals, and a way to deal with sparsely and unevenly distributed data. The degree to which these advantages are realized in practice is unclear, however. The ability of the eMLR method to recover the anthropogenic carbon signal is tested here using a global circulation and biogeochemistry model in which the true signal is known. Results show that regression model selection is particularly important when the observational network changes in time. When the observational network is fixed, the likelihood that co-located systematic misfits between the empirical model and the underlying, yet unknown, true model cancel is greater, improving eMLR results. Changing the observational network modifies how the spatio-temporal variance pattern is captured by the respective datasets, resulting in empirical models that are dynamically or regionally inconsistent, leading to systematic errors. In consequence, the use of regression formulae that change in time to represent systematically best-fit models at all times does not guarantee the best estimates of anthropogenic carbon change if the spatial distributions of the stations emphasize hydrographic features differently in time. Other factors, such as a balanced and representative station coverage, vertical continuity of the regression formulae consistent with the hydrographic context and resiliency of the spatial distribution of the residual field can be used to help guide model selection. The characteristic spatial scales of the modes of inter-annual to decadal variability in relation to the size of the North Atlantic, in concert with the station coverage available, place practical limits on the ability of eMLR to fully account for natural variability. Due to its statistical nature, eMLR only efficiently removes the natural variability whose spatial scales are smaller than the system analyzed.

1 Introduction

Since publication of the global oceanic cumulative mid-1990s anthropogenic carbon inventory estimate (Sabine et al., 2004), a measure of the time-integrated anthropogenic signal, attention has turned toward methodologies capable of monitoring spatio-temporal changes in that signal. Owing to the size of the oceanic carbon storage and the role of the ocean as a long-term sink of excess carbon dioxide, perturbations, progressive saturation or a decrease of the oceanic uptake rate (relative to expectations) can have large impacts on the atmospheric concentrations (Schuster and Watson, 2007; Corbiere et al., 2007; Le Quéré et al., 2007; Khatiwala et al., 2009). Accurate knowledge of the uptake rate and its inter-annual variability (McKinley et al., 2011) thus has important policy implications for carbon mitigation.

Independent assessments using atmospheric and oceanic carbon observations for the period 1995–2000 constrain the mean oceanic uptake rate of anthropogenic carbon to $2.2 \pm 0.3\,\mathrm{Pg\,C\,yr^{-1}}$ (Gruber et al., 2009). While estimates of the global uptake rate tend to converge (Wetzel et al., 2005; Takahashi et al, 2002; Mikaloff-Fletcher et al., 2006; Khatiwala et al., 2009; Takahashi et al., 2009), assessments diverge on a regional level, showing different uptake and storage patterns (Sabine et al., 2004; Waugh et al., 2006), especially in the Southern Ocean (Caldeira and Duffy, 2000; Lo Monaco et al., 2005a,b; Le Quéré et al., 2007). These differences have important mechanistic implications for the understanding and prediction of the marine carbon cycle and argue for improved observational estimates.

An accuracy target for the determination of the rate of change of anthropogenic carbon inventory of $0.1\,\mathrm{Pg\,C\,yr^{-1}}$ for each of the major ocean basins (3 Pg C globally over 10 yr, 10 % of the expected anthropogenic input for that period) was suggested in the Large Scale CO_2 Observing Plan (LSCOP) report (Bender et al., 2002) for the Repeat CO_2/Hydrography program. It is challenging to quantify the oceanic anthropogenic carbon concentration and its time rate of change, however. The first problem lies in the fact that anthropogenic carbon is usually defined as the difference between the contemporary dissolved inorganic carbon (DIC), i.e. the measured DIC, and an estimate of the natural DIC; that is, the DIC field thought to have existed in the absence of human activity (Gruber et al., 2009). The natural and anthropogenic carbon components are, however, indistinguishable from a measurement point of view. Separating them implies assumptions regarding the cycling of natural carbon. Another issues is that the anthropogenic carbon fraction is small relative to the background DIC concentration (of order $\leq 5\,\%$ of the DIC in the upper ocean). Even if the current analytical precision is sufficient to detect DIC changes on interannual to decadal time-scales (Brewer et al., 1997; Winn et al., 1998; Bates, 2001), natural variability confounds efforts to quantify the dynamics of the marine anthropogenic carbon sink on these scales (Keeling, 2005; Sabine et al., 2008; McKinley et al., 2011).

These difficulties are exacerbated by the limited number of data available and their spatio-temporal distribution. Basin or global-scale databases represent assemblages of data collected by individual cruises over many years. Owing to logistical limitations and since each cruise has its own scientific objectives, the large-scale spatio-temporal distribution of the data is not ideal. While new samples are often collected close to previously sampled stations, this is not always the case. As such, direct point-by-point data comparison in time may not be possible to infer changes on the basin scale if the intersecting datasets are too sparse. While a point-by-point analysis allows for a good control of the time difference between repeat samples locally (Levine et al., 2008; Sabine et al., 2008; Wanninkhof et al., 2010), this approach would only be applicable to a subset of the data for which repeat measurements are available. A strict section-by-section or station-by-station strategy would thus not be able to exploit the many samples for which no repeat exists. A form of extrapolation, which considers data in entire regions instead of constrained along sections, is thus desirable to make best use of available data.

Wallace (1995), Sonnerup et al. (2000) (in the context of ^{13}C) and Friis et al. (2005) proposed to compare empirical regression model representations of the measurements instead of directly comparing time-separated measurements to maximize data use, filter out the natural spatio-temporal variability and to generate spatial prediction. The Friis et al. (2005) implementation of this method is known as the extended Multiple Linear Regression (eMLR) approach. A few studies have described various aspects and limitations of the eMLR methodology either in models or applied to data (Sonnerup et al., 2000; Friis et al., 2005; Tanhua et al., 2007; Levine et al., 2008; Wanninkhof et al., 2010; Goodkin et al., 2011). We add to these previous efforts by addressing two points not thoroughly covered in the existing eMLR literature: the influence of regression model selection, and the effect of variable observational sampling networks on eMLR-derived estimates of the interannual to decadal change in anthropogenic carbon.

The eMLR procedure, under the constraint imposed by the number and locations of the available measurements, is here evaluated objectively using an ocean circulation model that includes carbon and nutrient biogeochemistry in which the true anthropogenic signal is known exactly. 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.

The principles of the eMLR theory are described first, using matrix notation to cast eMLR into the general framework of inverse problems. This is followed by a methodology section giving the details of the circulation and biogeochemistry model experiments used to generate the synthetic dataset on which the eMLR methodology is tested. The methodology section also includes a description of the calculations and of the mapping scheme used. Results are presented in three parts. The structure and variability of the anthropogenic carbon signal in the model are described first. Then, a summary of the regression results focusing on regression quality and formulae structure is given. The influence of various regression models and of changes in the observational network on the eMLR solutions is addressed in the last several sections. Basin-integrated inventory changes are discussed first, followed by layer-specific inventory changes and finally column inventory changes. A discussion of potential errors focusing particularly on the problem of inhomogenous data distribution in time and space precedes the conclusions.

2 eMLR theory

By design, regression models separate the fraction of the variance that can be explained by the model and the part that is due to noise. If suitable empirical regression models can be found to describe the DIC field in a spatial domain, and if it is assumed that the physical and biogeochemical processes acting in that domain are stationary and not affected by the anthropogenic perturbation, the noise (natural variability of DIC) can in principle be filtered out by regression and the anthropogenic signal revealed as the difference between model predictions of DIC at different times (Friis et al., 2005). Conceptually,

$$\Delta C_{\text{anth}}^{\text{eMLR}} = \mathbf{G}_2(D_2) - \mathbf{G}_1(D_2)\,, \tag{1}$$

where $\mathbf{G}_t$ are empirical model fits at times t derived from the respective datasets D_t. It is worth noting that a set of DIC predictions generated from a model fitted to one dataset but applied to the other dataset is necessary (in this example, $\mathbf{G}_1(D_2)$) to ensure that DIC predictions exist for all samples in the dataset D_2.

Tarantola (2005) gives the following expressions (his Eqs. 3.37 and 3.38) as possible forms of the least-squares estimator of the regression coefficients $\widetilde{\boldsymbol{c}}$ and the associated posterior covariance matrix $\widetilde{\mathbf{C}_\mathbf{c}}$ that constitute the regression model $\boldsymbol{Y} = \mathbf{Z}\cdot\widetilde{\boldsymbol{c}} + \boldsymbol{\epsilon}$:

$$\widetilde{\boldsymbol{c}} = \left(\mathbf{Z}^T\mathbf{C}_\mathbf{Y}^{-1}\mathbf{Z} + {\mathbf{C}_\mathbf{c}}^{-1}\right)^{-1}\left(\mathbf{Z}^T{\mathbf{C}_\mathbf{Y}}^{-1}\boldsymbol{Y} + {\mathbf{C}_\mathbf{c}}^{-1}\boldsymbol{c}_{\text{prior}}\right) \tag{2}$$

$$\widetilde{\mathbf{C}_\mathbf{c}} = \left(\mathbf{Z}^T\mathbf{C}_\mathbf{Y}^{-1}\mathbf{Z} + {\mathbf{C}_\mathbf{c}}^{-1}\right)^{-1}. \tag{3}$$

$\mathbf{C_Y}$ is the data covariance matrix and $\mathbf{C_c}$ is the prior covariance matrix of the estimator with mean prior densities given in the vector $\boldsymbol{c}_{\text{prior}}$. Exponents (T) and (-1) indicate the transpose and the inverse, respectively. Although this study uses noiseless synthetic data, a thorough treatment of these covariance matrices will be key for the application of eMLR with real data. This is, however, beyond the scope of this manuscript. $\mathbf{Z}$ is any design matrix containing the variables used as predictors and $\boldsymbol{Y}$ is a vector containing the DIC observations.

As indicated by Eq. (1), the eMLR estimate of anthropogenic carbon change is obtained by using two different sets of regression coefficients but only one set of data, resulting in estimates that are projected either forward or backward in time depending on the dataset used in the calculation. Using Eq. (2), the eMLR quantity that would be predicted with the dataset available at time t_2 is given by

$$\begin{aligned}\widetilde{\Delta C}_{\text{anth}|t_2}^{\text{eMLR}} &= \widetilde{\boldsymbol{Y}_{t_2}} - \widetilde{\boldsymbol{Y}_{t_1|t_2}} \\ &= \mathbf{Z}_{\mathbf{t_2}}\cdot\widetilde{\boldsymbol{c}_{t_2}} - \mathbf{Z}_{\mathbf{t_2|t_1}}\cdot\widetilde{\boldsymbol{c}_{t_1}} \\ &= \mathbf{Z}_{\mathbf{t_2}}\cdot\left[\left(\mathbf{Z}^T\mathbf{C}_\mathbf{Y}^{-1}\mathbf{Z}\right)^{-1}\left(\mathbf{Z}^T{\mathbf{C}_\mathbf{Y}}^{-1}\boldsymbol{Y}\right)\right]_{t_2} \\ &\quad -\mathbf{Z}_{\mathbf{t_2|t_1}}\cdot\left[\left(\mathbf{Z}^T\mathbf{C}_\mathbf{Y}^{-1}\mathbf{Z}\right)^{-1}\left(\mathbf{Z}^T{\mathbf{C}_\mathbf{Y}}^{-1}\boldsymbol{Y}\right)\right]_{t_1}\end{aligned} \tag{4}$$

in the limit of no available prior information ($\boldsymbol{c}_{\text{prior}} = 0$, ${\mathbf{C}_\mathbf{c}}^{-1} \to 0$) and with the "tilde" indicating empirical estimates. The subscripts t_2 and t_1 associated with the square brackets apply to every term in the brackets. $\mathbf{Z}_{\mathbf{t_2|t_1}}$ is the design matrix built from data at time t_2, but adjusted to utilize the variables included in the regression model derived from time t_1 ($\widetilde{\boldsymbol{c}_{t_1}}$). The notation $t_2|t_1$ is introduced to allow for different sets of predictor variables (i.e. different regression formulae) to be used in the derivation of the regression coefficients at either t_1 or t_2, a generalization of original eMLR (Friis et al., 2005).

Ideally, if the physical and biogeochemical processes that govern the spatial distribution of the tracers are stationary in time, the structure of best-fit regression formulae (i.e. the predictor variables used) should also be constant in time for a given region of the ocean. If the same set of predictor variables is used through time, i.e. $\mathbf{Z}_{\mathbf{t_2}} = \mathbf{Z}_{\mathbf{t_2|t_1}}$, and if the model is linear, Eq. (4) can be written as $\widetilde{\Delta C}_{\text{anth}|t_2}^{\text{eMLR}} = \mathbf{Z}_{\mathbf{t_2}}\cdot\left(\widetilde{\boldsymbol{c}_{t_2}} - \widetilde{\boldsymbol{c}_{t_1}}\right)$, which is the traditional form of eMLR (Friis et al., 2005).

In reality, as sampling intensity in different regions changes, $\mathbf{Z}_{\mathbf{t_1}}$ and $\mathbf{Z}_{\mathbf{t_2}}$ may not have the same number of rows (measurements) and these measurements may not be co-located geographically. As such, it is possible that the formulae of the regression models that minimize residuals in a region may change in time due to changes in the observational network, even without secular trends. Equation (4) explicitly accounts for this possibility. The degree to which changes in spatial sampling intensity affects the regression models and the degree of influence the form of the regression models ultimately have on the eMLR estimate is the subject of this study.

Note that the results can also be projected backwards in time, onto the data available at t_1: $\widetilde{\Delta C}_{\text{anth}|t_1}^{\text{eMLR}} = \mathbf{Z}_{\mathbf{t_1}}\cdot\left(\widetilde{\boldsymbol{c}_{t_2}} - \widetilde{\boldsymbol{c}_{t_1}}\right)$. If the numbers and locations of the measurements available change in time, maps produced by backward projecting the result at t_1 may differ from maps produced from forward projecting at t_2. eMLR can thus generate different results from the same data. The importance of this difference depends on sample coverage and mapping.

Equation (4) shows that predicted changes in the carbon concentration can occur as expected from differences in the vectors $\boldsymbol{Y}_{t_1}$ and $\boldsymbol{Y}_{t_2}$, but also from differences in the matrices $\mathbf{Z}_{\mathbf{t_1}}$ and $\mathbf{Z}_{\mathbf{t_2}}$ and from differences in the prior covariance matrices associated with variable $\boldsymbol{Y}$, $\mathbf{C}_{\mathbf{Y},\mathbf{t_1}}$ and $\mathbf{C}_{\mathbf{Y},\mathbf{t_2}}$. The measurement accuracy of DIC and alkalinity (Alk) have improved since the introduction of the certified reference material such that, for most samples taken during and after the World Ocean Circulation Experiment (WOCE), $\mathbf{C}_{\mathbf{Y},\mathbf{t_1}} \approx \mathbf{C}_{\mathbf{Y},\mathbf{t_2}}$. The measurement accuracy for DIC between cruises would vary by a factor of 2–5 prior to the introduction of reference material, such that changes in covariances can significantly contaminate the eMLR signal when using older datasets, as shown experimentally by Matear and McNeil (2003) and Tanhua et al. (2007). Equations (2) and (3)

do not formally consider errors associated with the predictor variables in **Z** but this can be achieved using a Monte-Carlo approach or more direct methods (Tarantola, 2005). Errors associated with the hydrochemical variables in **Z** are likely important in reality since no reference material is used for nutrient measurements and systematic biases are known to exist between measurements taken during different cruises.

Estimates of uncertainty around $\widetilde{\Delta C}^{\mathrm{eMLR}}_{\mathrm{anth}|t_2}$ can be obtained by linear propagation of the individual posterior uncertainties. Given that $\widetilde{\Delta C}^{\mathrm{eMLR}}_{\mathrm{anth}|t_2} = \widetilde{Y_{t_2}} - \widetilde{Y_{t_1|t_2}}$, and since the posterior covariance matrices $\widetilde{\mathbf{C_Y}}$ can be calculated from the design matrices and the posterior covariance matrix of the regression coefficients (Eq. 3) at each time by $\widetilde{\mathbf{C_Y}} = \mathbf{Z}\widetilde{\mathbf{C_c}}\mathbf{Z}^T$ (Tarantola, 2005), an estimate of precision for eMLR is

$$\widetilde{\sigma_{\mathrm{eMLR}}}^2 \approx \mathrm{diag}\left(\widetilde{\mathbf{C_{Y,t_2}}}\right) + \mathrm{diag}\left(\widetilde{\mathbf{C_{Y,t_1|t_2}}}\right) - 2 \cdot \mathrm{cov}\left(\widetilde{Y_{t_2}}, \widetilde{Y_{t_1|t_2}}\right), \quad (5)$$

with $\widetilde{\mathbf{C_{Y,t_1|t_2}}} = \mathbf{Z_{Y,t_1|t_2}} \cdot \widetilde{\mathbf{C_{c,t_1}}} \cdot \mathbf{Z_{Y,t_2|t_1}}^T$. By definition, the covariance term can also be expressed as $\mathrm{cov}(a,b) = \rho(a,b)\sqrt{\mathrm{Var}(a)\mathrm{Var}(b)}$, making the correlation (ρ) between co-located predictions of $\widetilde{Y_{t_2}}$ and $\widetilde{Y_{t_1|t_2}}$ explicit. This form of error propagation would be appropriate even if nonlinear regression models were considered since $\widetilde{\Delta C}^{\mathrm{eMLR}}_{\mathrm{anth}|t_2}$ is expressed as a difference between two terms, which is a linear operation.

Equation (5) shows that $\widetilde{\sigma_{\mathrm{eMLR}}}$ depends on the fit quality at time t_2 (the first term) and on an estimate of the regression precision achieved by applying the regression from t_1 with data from t_2 (second term). Finally, it importantly depends on the correlations between the prediction generated from each fit (the third term). If the estimates $\left(\widetilde{Y_{t_2}}\right)$ and $\widetilde{Y_{t_1|t_2}}$ are correlated, the overall estimated uncertainty decreases. On the other hand, if the predictions are uncorrelated the third term becomes small and the overall uncertainty around the eMLR result increases.

3 Methodology

3.1 Synthetic dataset and description of the model

A synthetic dataset with known anthropogenic carbon concentrations is used as a testbed. The synthetic dataset is constructed by sampling a global ocean circulation-biogeochemistry model (output provided by J. Dunne, Geophysical Fluid Dynamic Laboratory, NOAA, Princeton, NJ, USA) at the station coordinates given by the GLODAP (Key et al., 2004) and CLIVAR (defined operationally as data collected after GLODAP) datasets (Fig. 1) to reproduce the observed sampling grid. Our current working estimate of these datasets in the North Atlantic region represent 386 and 703 stations for GLODAP and CLIVAR, respectively. 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.

The analyses focus on 1995 and 2005. These years were chosen as they are representative of the modal sample density available for the GLODAP and CLIVAR datasets. Similarly, emphasis is given to July 1995 and July 2005 as July mimics the summer bias inherent in the original datasets (Key et al., 2004).

Our choice of the North Atlantic for this study is motivated by a number of factors. First, it is clearly a region of consequence for carbon uptake by the ocean (Sabine et al., 2004). Second, the complex hydrography and strong water mass variability in the North Atlantic pose particular challenges for empirically-based detection methods, as indicated by the global model-based eMLR results of Levine et al. (2008). Third, the relatively large number of measurements in this region suggests that it is an appropriate context within which to deconvolve uncertainties associated with the eMLR approach itself from uncertainties associated with the mapping process.

The simulator is composed of the NOAA/GFDL z-level coordinate Modular Ocean Model MOM4 general circulation model (Griffies et al., 2004, 2005; Gnanadesikan et al., 2006) and the Tracers in the Ocean with Allometric Zooplankton (TOPAZ) lower-trophic level biogeochemistry model (Dunne et al., 2005, 2007, 2008, 2010). Sea-ice dynamics are modeled by the GFDL Sea Ice Simulator (Winton, 2000).

The ocean model has 50 vertical layers and is resolved on a tripolar grid with an approximate resolution of 1°, improved to 1/3° meridionally near the equator. Synthetic profiles isolated at each station are not further sub-sampled in the vertical to mimic the observations, however. This results in a slight overestimation of the vertical sampling relative to the resolution of the data but the ocean is sufficiently well-sampled in the vertical. Horizontal interpolation errors are, for this problem, larger than vertical ones.

The TOPAZ biogeochemistry module is fully prognostic and includes all major nutrients (NO_3, PO_4, O_2, Si, DIC, Alk), labile and semi-labile dissolved organic matter pools, an iron cycle, ballasting of sinking particles, nutrient and light co-limitation, a microbial loop, three classes of phytoplankton and zooplankton. Details about the model formulation and performance are available in Dunne et al. (2010), Sarmiento et al. (2010) and Henson et al. (2009, 2010).

3.2 Simulation configurations and definition of anthropogenic carbon in the model

The model was initialized with World Ocean Atlas (2001) temperature, salinity and nutrients, GLODAP carbon and forced with the NCEP-derived CORE representation of atmospheric fields and fluxes (Large and Yeager, 2004, 2009; Griffies et al., 2009) over the period 1958–2006. Surface salinity was restored to observation with a relaxation time of 60 days.

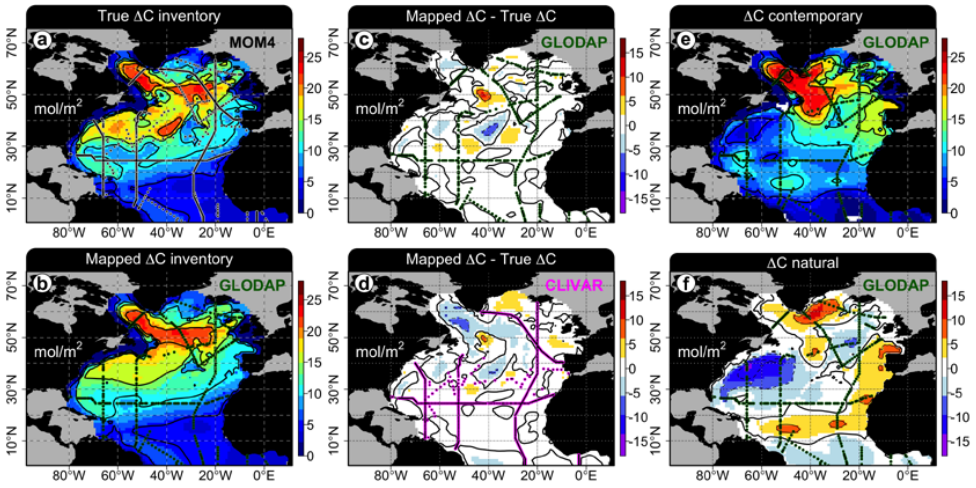

Fig. 1. (a) Change in anthropogenic carbon column inventory, in mol m^{-2}, between July 1995 and 2005 calculated on the original MOM4/TOPAZ grid. **(b)** Inventory change calculated after mapping the true values sampled at GLODAP stations. **(c)** Mapping error, difference between **(b)** and **(a)** for GLODAP. **(d)** Mapping error for CLIVAR. **(e)** Changes in contemporary and **(f)** natural carbon column inventories between July 1995 and 2005 mapped from GLODAP stations. Station locations are show in green (GLODAP) or magenta (CLIVAR). Both GLODAP and CLIVAR stations are plotted in **(a)**. In **(c)**, **(d)** and **(f)**, thin dashed (negative) and solid (positive) contour lines are drawn in increment of 6 mol m^{-2}. Thick contours mark 0 mol m^{-2}.

The strategy used to isolate the anthropogenic carbon concentration from the model is described by Rodgers et al. (2009). Briefly, the model was spun up for two repeating CORE cycles with fixed pre-industrial atmospheric CO_2 concentration after initialization. At this point, parallel integrations were performed: one with a prescribed atmospheric carbon dioxide transient boundary condition, yielding the contemporary carbon signal and one without, giving an estimate of what the evolution of the natural carbon would have been had the atmosphere remained stable at pre-industrial pCO_2 levels. These parallel simulations were repeated for five additional CORE cycles with the atmospheric CO_2 concentration increasing monotonically throughout the five cycles as prescribed by the known evolution of historical atmospheric pCO_2. The last cycle is used as a model surrogate for years 1958–2006 and provides the basis for this work. Since both branches of integration were forced with exactly the same forcing fields, the physical state variables are identical and the only difference between the two runs are the concentrations of carbon dioxide in the oceanic and atmospheric reservoirs. The anthropogenic carbon concentration is operationally defined to be the difference between the two runs. The model global anthropogenic carbon inventory in 1995 is 104.9 Pg C, a value within errors of the observational estimate (Sabine et al., 2004).

3.3 Regressions, statistics and eMLR calculations

First-order additive linear models were fitted to the synthetic DIC datasets extracted from the monthly mean fields of the MOM4/TOPAZ simulations in 1995 and 2005 sampled at both GLODAP and CLIVAR station locations. All 255 possible models, from single-term to 8-term models, were considered, using the following set of oceanographic variables (salinity, potential temperature, nitrate, phosphate, silicate, apparent oxygen utilization, oxygen, salinity): $Z \subseteq \{S, \theta, NO_3, PO_4, Si, AOU, O_2, Alk\}$. An offset term (i.e. y-intercept) is implicitly included in each fit but this term is not included in the following discussion for simplicity. See Table S1 in the supplementary material for a list of the model formulae.

The best regression models chosen from all possible first-order models were identified for each size class (1 to 8 term models) and across all size classes and for each horizontal layer and each month from January to December for the nominal years 1995 and 2005 to investigate the effect of temporal, physical and biological variability on the ability of simple linear regression models to fit oceanographic data. The minimum Akaike Information Criterion (AIC) was used as a guide for model selection across the complexity spectrum. AIC addresses the bias-variance trade-off problem when comparing models of different complexity and minimizes the risk of over-fitting. AIC is defined as $\mathrm{AIC} = -2\ln(L) + 2k$, where L is the maximum likelihood of the fitted model and k is the number of parameters in the model. AIC is then simply a measure of the residual sum of squares misfit (L) with a penalty added ($2k$) that is a function of the number of terms in the model (Burnham and Anderson, 1998). For a given set of data, models corresponding to the smallest AIC values represent the best consensus between fit quality and model complexity.

To tease apart the influence of changes in the observational network from regression model selection on the eMLR results, the following cases are considered. First, realistic calculations are made where regression models are derived from GLODAP data in 1995 and CLIVAR data in 2005. Such "hybrid" results projected both backwards in time on the GLODAP data and forward in time onto the CLIVAR stations are considered:

$$\Delta C^{\text{hybrid}}_{\text{GLODAP}} = \mathbf{G}^{2005}_{\text{CLIVAR}}(D^{1995}_{\text{GLODAP}}) - \mathbf{G}^{1995}_{\text{GLODAP}}(D^{1995}_{\text{GLODAP}}) \quad (6)$$

$$\Delta C^{\text{hybrid}}_{\text{CLIVAR}} = \mathbf{G}^{2005}_{\text{CLIVAR}}(D^{2005}_{\text{CLIVAR}}) - \mathbf{G}^{1995}_{\text{GLODAP}}(D^{2005}_{\text{CLIVAR}}) \,. \quad (7)$$

The hybrid results are contrasted with idealized calculations where the observational networks are held fixed in time. Two scenarios are considered, one for each set of stations:

$$\Delta C^{\text{fixed}}_{\text{GLODAP}} = \mathbf{G}^{2005}_{\text{GLODAP}}(D^{1995}_{\text{GLODAP}}) - \mathbf{G}^{1995}_{\text{GLODAP}}(D^{1995}_{\text{GLODAP}}) \quad (8)$$

$$\Delta C^{\text{fixed}}_{\text{CLIVAR}} = \mathbf{G}^{2005}_{\text{CLIVAR}}(D^{2005}_{\text{CLIVAR}}) - \mathbf{G}^{1995}_{\text{CLIVAR}}(D^{2005}_{\text{CLIVAR}}) \,. \quad (9)$$

3.4 Mapping

Mapping, that is the horizontal extrapolation of point samples to a basin-scale grid, is a necessary step in calculating inventories from the eMLR predictions as these are produced only at the GLODAP or CLIVAR stations. Mapping was performed using a fixed exponential covariance function with a longitudinal correlation scale of 15.5° and a latitudinal scale of 7.4° above 3500 m, or 7.4° for both scales below that depth. Analysis of the semi-variograms, experimentation with the length-scales and other kriging control parameters showed these scales to be appropriate. This scheme was chosen to mimic the objective mapping process used by Key et al. (2004) who used typical length scales of 1550 and 740 km above 3500 m and 740 km in both direction below that depth, and to ease the computational burden. In light of the thousands of maps that were produced, a fully adaptable kriging scheme for each map was not practical. Inventories were calculated from fields mapped to a regular $1° \times 1°$ grid.

4 Changes in DIC distribution

A description of the target signal (change in anthropogenic carbon) and its components (change in natural and contemporary carbon) is provided first, before the eMLR results, to provide context for the signal in relation to the variability captured by the model and the sampling network.

4.1 The "true" target signal

Figure 1a shows the modeled change in column inventory of anthropogenic carbon between July 1995 and July 2005. Figure 1a represents the target signal that eMLR aims to recover. Figure 1a is calculated on the original model grid after subtracting the control (natural carbon) from the transient (contemporary carbon) component. Figure 1a shows that regions with large inventory changes associate closely with water mass formation regions that are also high uptake regions, notably the Labrador Sea Water and the North Atlantic Subtropical Mode Water formation regions, but also reflect water mass reorganization, gyre wobble and frontal shifts in the control simulation.

Both the GLODAP and the CLIVAR observational networks are overlain as a series of dots on Fig. 1a, showing how some notable high-change regions are entirely missed by the sampling. One such high-change feature, with column inventory differences above 20 mol m^{-2} and centered around 35° W–35° N, is missed entirely by both the GLODAP or the CLIVAR stations. Another localized high-change feature is situated near 60° W–38° N and is similarly omitted in the respective datasets. The Labrador Sea is currently only sampled by the GLODAP stations in our data compilation (post-GLODAP data in this region will soon become available).

Figure 1b shows the vertically integrated anthropogenic carbon inventories resulting from sampling the model at the GLODAP station locations and extrapolating horizontally to the basin scale using the mapping method described previously. The mapping was performed separately for each model level. Mapping, using either the GLODAP or CLIVAR station distribution, results in a slight overestimation of the vertically integrated signal in the subtropics and in underestimation in the subtropical/subpolar transition and in the Labrador Sea (Fig. 1c and d). Vertically integrated biases resulting from mapping are most significant in the unsampled regions East of the Grand Banks and in the central North Atlantic (40° N, 40° W). In these restricted areas, mapping errors can be as much as half the size of the anthropogenic signal (about ± 10 mol m^{-2} in absolute terms). These unsampled regions are also zones experiencing the highest magnitude of temporal carbon variability in the North Atlantic in the simulations (Rodgers et al., 2009). When integrated over the basin, mapping errors are smaller than other sources of uncertainties, however. On each horizontal layers, the kriging uncertainty (uncertainty around the central kriging estimator) resulting from mapping error-free values sampled at GLODAP or CLIVAR stations is typically of similar magnitude as the absolute error (difference between the true value and the central kriging estimator). Propagation of the mapping uncertainty is not considered in the following discussion, where only the central kriging estimator is used as a diagnostic.

4.2 Changes in simulated contemporary and natural carbon distributions

Figure 1e and f show the vertically integrated DIC column inventory changes in the transient and control simulations (the

two components used to calculate the anthropogenic signal), mapped from the set of samples taken at GLODAP locations. The change in the column inventory of contemporary carbon for the transient simulation between July 1995 and July 2005 (Fig. 1e, $\Delta C_{\text{contemporary}}$) reveals substantial carbon accumulation in the subpolar gyre region, the European Basin and at the southern edge of the subtropical gyre ($\approx 15^\circ$ 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 for by the changes over the same decadal interval in the control simulation (Fig. 1f, $\Delta C_{\text{natural}}$), highlighting the importance of natural variability. For example, the Western Subtropical Atlantic shows a drastic decrease in natural carbon between 1995 and 2005, which, when added to the transient run, results in substantial carbon uptake in the subtropical mode water formation region (Fig. 1a, b), consistent with what is expected from previous studies (Bates et al., 1996; Lee et al., 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 (Fig. 1f).

Inspection of horizontal maps of DIC change in the control simulation between 1995 and 2005 suggest that the systematic negative change in vertical inventory (Fig. 1f) in the North American Basin is caused primarily by a decrease in the DIC concentrations (> 5–$10\,\mu\text{mol}\,\text{kg}^{-1}$) in the deep model ocean (> 2200 m). These deep DIC changes are accompanied by a decrease in the concentration of the other nutrients, an increase in oxygen, and a slight warming. The Labrador Sea and subpolar basin show large increases in carbon and in nutrient concentrations, a decrease in oxygen concentrations and strong increases in salinity and potential temperature. These changes are topographically constrained to the west of the Mid-Atlantic ridge below 3000 m, but the changes between 2200 and 3000 m suffice to explain the drop in column inventory visible in the northeastern Atlantic (25° W, 50° N, Fig. 1f). These patterns indicate that variability in the convective activity and export of the Labrador Sea and downstream adjustments of the Deep Western Boundary Current and interior properties are responsible for the large-scale column inventory changes in the northern and western Atlantic (Fig. 1f). The increase in column inventory simulated by the control run at the southern edge of the subtropical gyre and eastern Atlantic is due to gyre dynamics. Increases in the DIC field are observed in this region between 150 and 700 m, along with changes in other tracers. These changes are consistent with a northward contraction of the subtropical gyre.

Differences in the simulated annual mean sea surface height (SSH) between 1995 and 2005 agree with the interpretation given above (not shown). The patterns of change in SSH do not reflect the North Atlantic Basin drop in carbon inventory seen in Fig. 1f, suggesting the source of that feature is in the deep ocean. On the other hand, SSH varies consistently with the signal observed at the eastern and southern edge of the gyre (Fig. 1f). The regions with increasing carbon in the control simulation (Fig. 1f) coincide with the regions of highest interannual variability identified by Cromwell (2006) from an analysis of satellite SSH data in the North Atlantic. The source of the positive deviation of the carbon inventory in the subequatorial (10–20° N) and eastern North Atlantic is in the upper few hundred meters. This pattern likely reflects a real mode of interannual variability captured by the model.

The subtropical region with strong negative change in the column carbon inventory (Fig. 1f) is identified as a low SSH variability region by Cromwell (2006). This is further evidence that the strong and coherent signal of Fig. 1f is not due to interannual variability in the upper thermocline. This signal is rather associated with the Labrador Sea Water and is consistent with the observational analysis of Curry et al. (1998) who reported how deep subpolar perturbations caused by changing convection in the Labrador Sea propagate to the subtropics. These patterns of DIC inventory changes simulated by the model are qualitatively consistent with important known patterns of SSH variability over the North Atlantic that also affect field observations.

5 Model selection and variability of regression performance

Oceanographic applications of eMLR have typically relied on one of two approaches to address the issue of regression model selection. On one hand, models are chosen a priori based on knowledge of the physical and biogeochemical processes or data availability. On the other hand, the model selection problem is addressed statistically, relying on stepwise linear regression. In this section, we explore the ability of various regression formulae to explain the data as a function of depth and time and explore the spatio-temporal continuity of the statistically selected models. The analysis shows that, for the most part, there is convergence of statistically selected model formulae across multiple depth intervals. This is consistent with the fact that water mass differences are responsible for most of the variance along the horizontal layers in the domain analyzed. Best-fitting model formulae change in time, however, being affected strongly by variations in the sampling network.

5.1 Regression formulae

Figures 2 and 3 show the vertical continuity of the statistically selected best-fit model structures for each complexity class and overall for the July 1995 GLODAP or the July 2005 CLIVAR datasets. In these figures, the horizontal axis represents regression model number (see Table S1 for model definitions). These models represents the full suite of possible permutations for eight predictor variables, beginning with a model with only one term (model 1) to the model

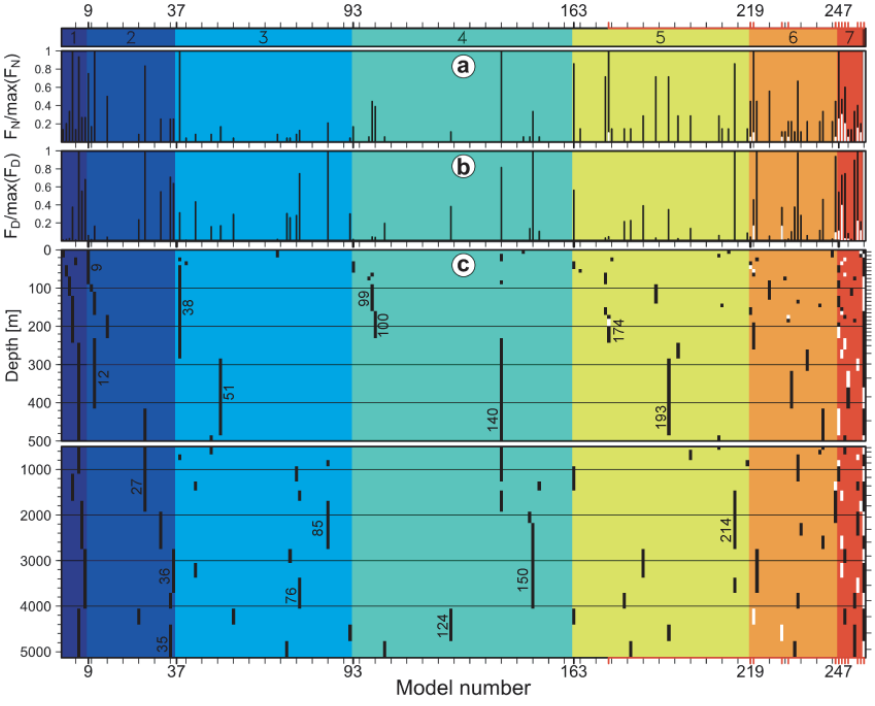

Fig. 2. Summary of the best fitting linear models for the July 1995 GLODAP synoptic synthetic dataset. Background colors identify models size classes (1 to 8). **(a)** Relative frequency (F_N/max(F_N)) with which models are selected in each size class (minimum root-mean-square error, black bars) and overall (minimum AIC, white bars). Frequency is computed based on the number of model layers (F_N) normalized to the most frequently identified model (max(F_N)). **(b)** Same as **(a)** but for frequency weighted by the thickness of each layer (F_D/max(F_D)). **(c)** Models with with lowest AIC in each size class (black bars) and overall (white bars, red ticks on top and bottom x-axes) and each depth layer. Tick marks on the right show boundaries between model layers. Tick marks on top and bottom show model number (in steps of 5). The first model number of each size class is indicated, except for size classes 1 and 8 (number 1 and 255).

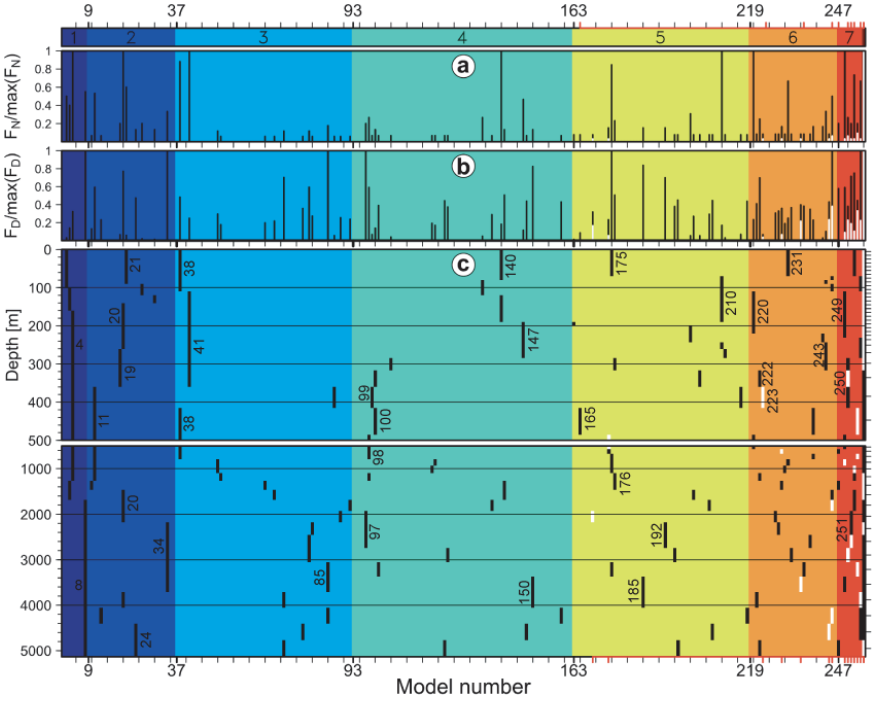

Fig. 3. Same as for Fig. 2 but using the July 2005 CLIVAR synoptic synthetic dataset.

containing all eight terms (model 255). The color strip at the top that matches the figure background summarizes information about model complexity with each color corresponding to increments in the total number of predictor variables. Panel c in these figures indicates the models that are statistically best in each size class (black vertical segments) or overall (white vertical segments), as a function of depth. Panels a and b summarize the frequency with which particular models are selected throughout the water column, plotted either as a number frequency (panel a) or weighted as a function of layer thickness (panel b).

Parallel analyses for the complementary July 2005 GLODAP and July 1995 CLIVAR cases indicate that changing the sampling networks influences the model selection process more than interannual variability does for a constant set of stations. Given a constant sampling network, only few temporal changes in the statistically optimal formula structure are detected (over all models or within model complexity classes) and these typically only involve one of the terms in the formula. These term swaps are also consistent with the vertical patterns of changes in standard deviation seen between datasets constructed from the 1995 and 2005 sampling of the model fields (Appendix A).

A set of regression predictors optimized from data taken in a particular depth range may not necessarily represent the best set on a different depth layer (Figs. 2 and 3). This is because the processes governing the distribution of tracers vary with depth. Similarly, a model derived from a particular set of stations on a given layer may not be suited to a different subset of stations on the same layer if the two sampling networks are sparse relative to the main variance pattern characteristic of the particular layer. In this latter case, it is not necessarily because processes governing the variance on the layer have changed, but because the sampling networks capture the variance pattern differently.

As Figs. 2 and 3 show, the set of regression models selected by the GLODAP or CLIVAR observational networks differs in each complexity class. These observational networks emphasize various hydrographic structures differently owing to the presence, absence, and density of sampling stations in certain areas. The sampling density in CLIVAR emphasizes the Eastern Atlantic and the subtropical gyre. GLODAP, in spite of having fewer stations, samples the North Atlantic more homogeneously with stations in the Irminger Sea, the Iceland Basin and the Labrador Sea, giving relatively more weight to the subpolar region than CLIVAR does. These regions are characterized by anomalously low temperature and low salinities relative to the basin average. As a result of the differences in sampling, regressions derived from the North Atlantic GLODAP data may a priori be considered more representative of the mean basin-scale while the CLIVAR fits may be more influenced by the subtropics and the subtropics/subpolar transition.

A quantitative analysis of the terms in the selected formulae in Figs. 2 and 3 as a function of depth reflects the differences in network representativeness (Fig. 4). For instance, analysis of the statistically selected formulae highlights the importance of salinity in the top 300 m as an explanatory variable in the regressions derived from the GLODAP dataset (Fig. 4a, b). In contrast, temperature and oxygen replace salinity in many of the formulae produced from the CLIVAR stations in this depth range (Fig. 4b–f). This is because the dominant source of variance in the CLIVAR set mostly represents the subpolar to subtropical contrast and is less influenced by extreme regional features such as the East/West Greenland Current and the Labrador Sea than GLODAP. Salinity takes a relatively more important role in CLIVAR between 400 to 1200 m (Fig. 4c, d). This reflects the influence of the Mediterranean Sea Overflow water in the Eastern Atlantic, which is relatively more frequently sampled in CLIVAR. Silicate is more frequently present in the formulae in that depth range in the regressions derived from the GLODAP set of samples (Fig. 4a, b). Common features also exist, however, between the formulae structures generated by the two sampling grids (Fig. 4e, f). For instance, the role of phosphate at intermediate depths (200–1500 m) is clear for both networks. Similarly, alkalinity is recurrently selected in the deep ocean (below 2000 m). This is due to the observed longitudinal difference in alkalinity across the mid-Atlantic Ridge. Overall, nitrate and AOU are the variables selected least often in the formulae. This may be because denitrification and nitrogen fixation influence the nitrate distribution strongly, but only weakly impacts the large-scale DIC gradients, resulting in phosphate being the preferred variable for the purpose of fitting basin-scale DIC patterns. Similarly, there is an assumption of saturation in the calculation of AOU that may explain why AOU is a slightly more incompatible variable than other predictors in linear regression models of DIC.

5.2 Regression quality

In addition to model structure, regression quality also varies with depth. Overall, the quality of the best fits, as measured by the AIC values, is lower towards the top than towards the bottom (Fig. 5a, c). Many models possess a thin layer centered around 1500 m where fit quality is better than in the layers just below (2000 m, Fig. 5a, c). Given that the vertical profile of the range of AIC values across all 255 models on each layer show a maximum between 800 and 1500 m (Fig. 5b), model selection can make a significant difference in this layer whose variance is dominated by the contrast between the extreme properties of Mediterranean Sea Overflow Water, Subpolar Mode Waters and Labrador Sea Water, water masses that are separated by a sharp front located along the North Atlantic Current. On the other hand, the vertical profile of the layer-specific AIC range across all models shows a minimum between 2000 and 2700 m suggesting that this layer is a priori less sensitive to the form of the particular regression formula. This does not mean that the fits in this in-

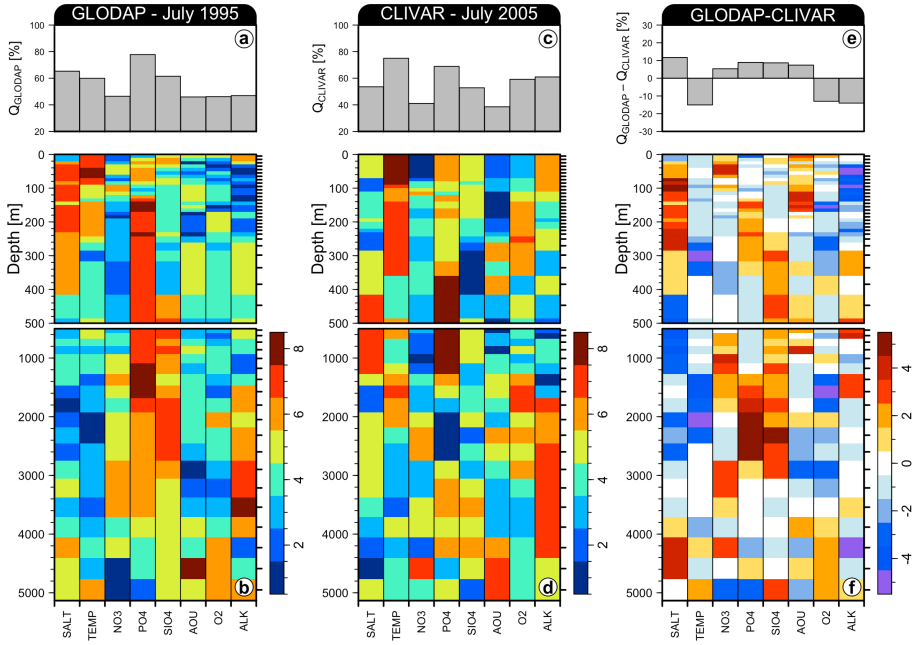

Fig. 4. Summary of the frequency of occurrence of the variables in the formulae of the best fitting models in each size class given the 1995 GLODAP **(a, b)** or the 2005 CLIVAR **(c, d)** stations and the difference between the two cases **(e, f)**. The color scale indicates the total number of times a variable is present summed over each best-fit formula across all size classes for each horizontal layer (v_i, the maximum is 8 for each layer). **(a, c, e)** Relative frequency of occurrence of each variable integrated over all depth layers and normalized to the maximum possible occurrence, for each variable j, $Q_j = \frac{\sum_{i=0}^{D} v_i}{8D}$, where D is the number of vertical layers.

terval are necessarily good, however. The minimum AIC profile shows a relative maximum between 2000 and 2700. The difference between the maximum and minimum AIC value is lowest below 4000 m where the fits are also best, suggesting a priori that many equivalent models can be used to fit the DIC field in that range.

While Figs. 2 and 3 indicate that there is some volatility in terms of the best-models identified as a function of depth, quite a few models have AIC values within 10 % of the depth-specific AIC range from the minimum AIC in each layer (highlighted in black, Fig. 5a). Differences in AIC values can be relatively small between many of the regression models. Often, these closely fitting formulae fall in related groups, e.g. the nitrate term replaces the phosphate term, oxygen and AOU swap. While a strict identification of the minimum AIC values overall or within complexity classes can result in model formulae with different structures, summary regression statistics like AIC suggest that the DIC field can be fitted to similar degrees of precision using a variety of different models. Although this work does not consider measurement uncertainty, this additional source of noise would further blur boundaries between regression formulae. These considerations about model fit suggest the possibility of using closely related models when convenience dictates, for example to maximize data coverage in cases when measurements for particular tracers are missing. These results also indicate, though, that the importance of regression model selection varies with depth, with some layers being particularly sensitive to the choice of regression formula.

eMLR relies on differences between fit predictions and not on single fit quality, however. This implies that co-located systematic misfit errors, that is the systematic error of using the wrong empirical model to represent the true, yet unknown, underlying model governing the distribution of the anthropogenic carbon fraction at a given time, can cancel during subtraction of the model predictions if the misfit error is similar at both times (Goodkin et al., 2011). This systematic misfit cancelation effect reduces the influence of misfit error on the final carbon change estimates and can attenuate the role of regression model selection on the overall eMLR results.

The systematic misfit cancelation effect is greatest when regression formulae and the sampling grid are temporally invariant, and when the magnitude of the spatial variance pattern that drives the regression is not greatly influenced by temporal variability. In the more realistic case of a non-homogenous and temporally variable sampling network, the systematic misfit cancelation effect is further regulated by the fact that regression fits not only reflect temporal changes but also changes in how the main spatial variance is captured by the various sampling networks (see Appendix A). This means that the geographical distribution of the regression residuals may change, thus modulating the net influence of the systematic misfit cancelation effect and the importance of regression

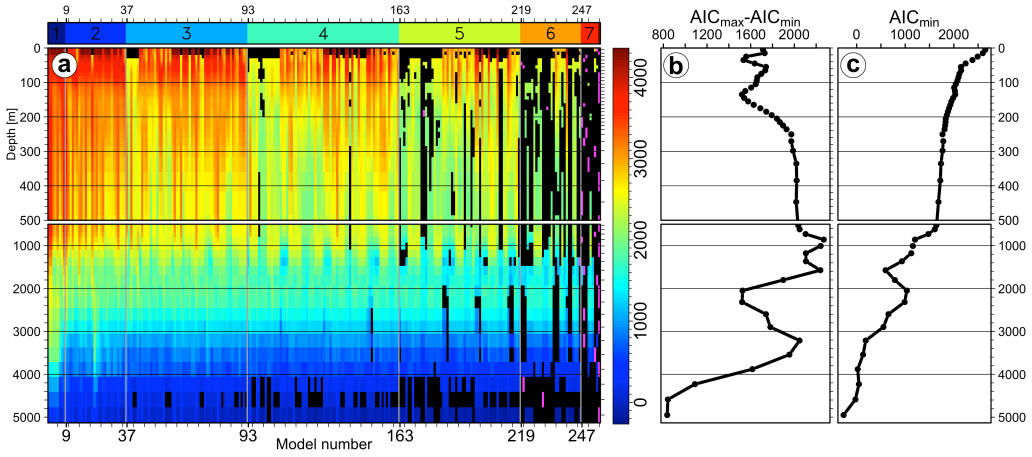

Fig. 5. (a) AIC values as a function of model number (strategy 2) and depth for the July GLODAP 1995 dataset. All models with AIC values within 10 % of the depth-specific range in AIC of the minimum AIC value at each depth (highlighted in magenta) are highlighted in black. Tick marks on the right show the vertical location of model layers. Corresponding vertical profiles of **(b)** the depth-specific range in AIC and **(c)** the minimum AIC values.

model selection on the eMLR results. This can happen even when the same set of predictor variables are used to fit both sets of stations and even if the overall quality of the regression fits, as measured by summary diagnostics like AIC for example, are very good. This is because the empirical definition of what is considered noise and what is considered signal, even for a fixed formula, may change depending on the distribution of stations.

Since fewer outliers with smaller misfit are generally found for regression fits from more complex models, one could expect that the influence of the systematic misfit cancelation effect is less when summary regression diagnostics are indicative of good fit, such as when AIC or the residual sum of squares is low, as opposed to when summary diagnostics of regression fit quality are poor. Clearly, if fit residuals are small, their addition or subtraction will have a smaller effect on the net results but that also means that more of the variance in DIC will be considered to be part of the anthropogenic signal and less will be considered to be associated with natural variability. This raises the question of model overfitting. When the observational network varies, it may be beneficial to accept worse local fit that make fewer assumptions about the underlying structure of the anthropogenic signal relative to the noise and are less susceptible to interpolation errors than to select regressions that fit observations better but that may be overly specialized towards local features and whose signal/noise partitioning will be strongly affected by small changes in the observational network. When the sampling grid is changing in time, it is possible that a regression model becomes highly specialized for one sampling network while being unrepresentative of another set of stations or of the main variance pattern on the layer, even when the two station subsets are taken at the same time and from the same general geographic domain. This may lead to interpolation errors that directly impact the eMLR estimates. Model selection, in the context of eMLR, should then be mindful of the spatio-temporal scales characteristic of the domain analyzed in relation to the overall objectives of the study (regional evaluations, basin-scale integrals, etc). Selecting regression simply because they lead to smaller residuals can be helpful, but this his not a sufficient criterion for model selection.

6 Recovery of the change in anthropogenic carbon signal by eMLR

While the question of variable station coverage and associated dataset variance does not arise when dealing with exactly repeated datasets, as in previous model-based eMLR evaluation studies (Levine et al., 2008; Goodkin et al., 2011), it is an important issue for more realistic basin-scale eMLR application. Previous applications of eMLR have required the structure of the regression formula to be constant as a function of time and derived the anthropogenic signal by difference between the regression coefficients. However, direct subtraction of the regression coefficients is only possible because the models are linear. The equivalent signal can also be obtained by subtracting the predicted DIC values obtained after parallel application of the regression equations to the data from both time points (Eq. 4). This second approach opens the conceptual possibility of using separate regression models, possibly nonlinear models, derived independently at each time point.

The main argument for using a constant model structure in time is that the physical and biogeochemical processes maintaining the DIC field are relatively constant and should thus be constrained by the same empirical models. In practice, there is no guarantee that empirical formulae represent these

physical and biogeochemical processes accurately. In addition, if the observational network varies, the variance in the data will change and regression formulae will match these different patterns of variance, such that the concept of "best" formula becomes an ad hoc function of depth, stations distribution and sampling density.

This section investigates the overall performance of eMLR and contrasts results obtained by the two limiting conceptual approaches described above, namely: strategy (1) use of a composite of statistically optimized formulae with sets of explanatory variables that are allowed to vary in time and as a function of depth, and strategy (2) use of regression formulae with a constant set of explanatory variables at all depths and times but with regression coefficients optimized independently at each depth and time. Furthermore, to highlight the role of changing the observational network, the change in carbon resulting from these two strategies using GLODAP data in 1995 and CLIVAR data in 2005 (ΔC^{hybrid}) are contrasted with complementary analyses that hold the observational networks fixed in time (1995 GLODAP compared to 2005 GLODAP, and 1995 CLIVAR compared to 2005 CLIVAR, ΔC^{fixed}).

The three-dimensional North Atlantic eMLR results and the associated absolute errors are presented below at various levels of integrations. Basin-integrated inventory changes are presented first. Because basin-scale inventories integrate over the whole volume, they are less sensitive to random errors. Vertical profiles of the layer-specific inventory changes are presented next. These integrate horizontally, along the direction along which the regressions are performed. Column-inventory changes are presented last. These represent vertically integrated results, perpendicular to the direction along which the eMLR analysis is performed.

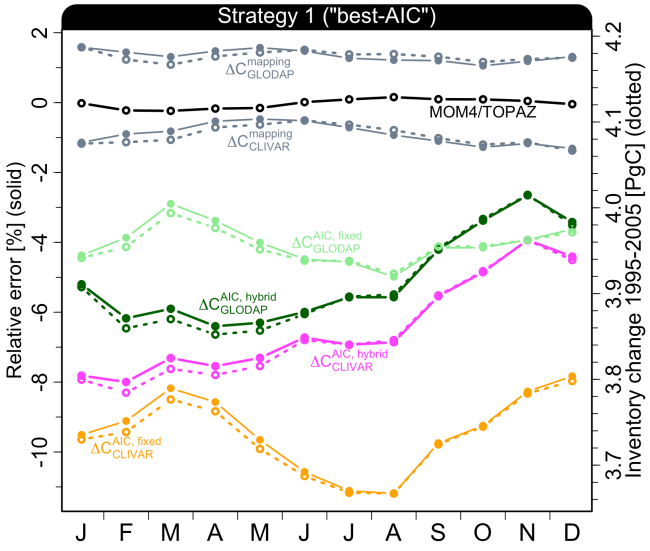

Fig. 6. Relative error (left y-axis) and absolute value (right y-axis) of the anthropogenic carbon inventory change calculated month-by-month between 1995 and 2005. (MOM4/TOPAZ, black) Inventory changes calculated from the "true" values on the original model grid, ($\Delta C_X^{\text{mapping}}$, gray) mapping the "true" values sampled at the GLODAP or CLIVAR stations, ($\Delta C_X^{\text{hybrid}}$, dark green and magenta) after mapping the hybrid composite best-AIC solutions obtained from regressions specific for each time and sampling network projected either onto the GLODAP or CLIVAR stations, or ($\Delta C_X^{\text{fixed}}$, light green and orange) by holding the observational networks fixed in time while allowing for the fixed-network best-AIC models to vary in time.

6.1 Basin-scale inventories

6.1.1 "Best AIC" strategy

The simulated (true) change in North Atlantic carbon inventory between 1995 to 2005 is 4.12 Pg C (MOM4/TOPAZ) and does not vary much throughout the year indicating that seasonality is fairly constant between these two years (Fig. 6, the minimum is 4.11 Pg C in March and the maximum is 4.13 Pg C in August; differences are calculated month-by-month, i.e. January 2005–January 1995, etc.). In contrast, the relative errors of the basin-integrated "best AIC" eMLR estimates (strategy 1) systematically underestimate the true values and vary seasonally from about −4 % in November to −8 % in February when the observational network is allowed to change for results projected onto the CLIVAR stations ($\Delta C_{\text{CLIVAR}}^{\text{AIC, hybrid}}$), or from about −3 % to −6 % for GLODAP ($\Delta C_{\text{GLODAP}}^{\text{AIC, hybrid}}$, Fig. 6).

Mapping explains some of the offset between the GLODAP and CLIVAR results: the underestimation is less severe in the GLODAP case than in the CLIVAR case because the mapping process introduces a compensating ≈1.5 % overestimation in GLODAP but ≈1 % underestimation in CLIVAR, as measured by maps produced from true values sampled at the station locations. Even when the observational networks are held fixed in time, mapping can only account for about half the difference between the GLODAP and CLIVAR cases (difference between $\Delta C_{\text{GLODAP}}^{\text{AIC, fixed}}$ and $\Delta C_{\text{CLIVAR}}^{\text{AIC, fixed}}$, Fig. 6), however. The other half represents a systematic bias that results from the influence on regression of the different coverage of the sampling networks. The fact that errors associated with $\Delta C_{\text{CLIVAR}}^{\text{AIC, fixed}}$ are about −10 % while $\Delta C_{\text{GLODAP}}^{\text{AIC, fixed}}$ errors are about −5 % gives further support to the idea that the GLODAP station coverage is more representative of the overall North Atlantic domain than CLIVAR, even if GLODAP has fewer stations. Because the hybrid inventory changes (with temporally varying sampling networks) represent the convoluted influence on the signal/noise partitioning due to both temporal variability, which is a common factor for both networks, and changes in observational networks, the hybrid results tend to be intermediate between the two limiting fixed-network cases.

The seasonal signals for $\Delta C_X^{\text{AIC, fixed}}$ and $\Delta C_X^{\text{AIC, hybrid}}$ are amplified relative to the ideal mapping-only results

($\Delta C_X^{\text{mapping}}$, Fig. 6, X is a dummy variable to indicate GLODAP and CLIVAR). Since the seasonal amplitude of the mapping error is small, the amplified seasonal cycle of the eMLR estimates must reflect seasonal variations in the systematic misfit cancelation effect. Seasonal differences in basin-integrated inventory amount to about 4 % when the sampling network changes. The GLODAP-CLIVAR differences are less, about 2–3 %, in the fixed network case. The seasonal amplitude for CLIVAR is greater than for GLODAP in the fixed-network case, but the shape of the seasonal cycle is otherwise mostly parallel between the GLODAP and CLIVAR results. Nevertheless, the shape of the seasonal cycle is different between the hybrid and fixed-network results (Fig. 6).

Standard summary diagnostics of overall regression quality (AIC, mean residual sum of squares, coefficient of determination) are all indicative of excellent fits for both sampling networks, at all times and depths for the "best AIC" fits. All "best AIC" fits are significant at $p \ll 0.001$ with $R^2 > 0.98$ and with mean standard error of the residuals typically smaller than the modern measurement uncertainty of DIC ($\approx 4\,\mu\text{mol kg}^{-1}$, consistently smaller than $2\,\mu\text{mol kg}^{-1}$ below 500 m). Upon closer inspection, small seasonal variations in fit diagnostics exist (mostly in the top ≈ 200 m) indicating that for both GLODAP and CLIVAR and both in 1995 and 2005 it is more difficult to fit first-order linear models to summer and fall data (May–October) than to winter data (January–March). The relative error of eMLR-derived basin-scale inventory changes is larger in summer and smaller in winter for the $\Delta C_X^{\text{AIC, fixed}}$ cases, following broadly the seasonal cycle in fit quality. It is not clear, however, how to relate the seasonal changes in fit quality to the net effect on the $\Delta C_X^{\text{AIC, hybrid}}$ results as the seasonal cycle of fit quality diagnostics and the hybrid inventory errors are phase shifted. The errors associated with the basin-integrated $\Delta C_X^{\text{AIC, hybrid}}$ inventory change estimate are smallest in the fall, early winter and largest in the spring.

Deep convection in winter, shifting of the Gulf Stream and North Atlantic Current, shoaling of the mixed layer in spring and blooms all contribute to the presence of sharp horizontal property gradients across the basin that are difficult to properly represent with linear models empirically defined over broad geographic scales from sparse datasets. Owing to the systematic misfit cancelation effect, however, it is not only these processes and features that matter, but also how each of these processes and features vary inter-annually and how each affect the datasets used at each time point as a whole. For example, convective activity may be a critical process locally, but if the sampling network (and the regression fit) is not greatly influenced by the convective region, this will only have a small influence on fit quality. On the other hand, if a second sampling network is influenced by the convective region (arguably GLODAP) and it is compared with a fit from another network that is not (arguably CLIVAR), changes in convective activity will have a larger effect on the overall eMLR estimate as the systematic misfit cancelation effect will not be able to correct the bias introduced by the dynamical mismatch between the two empirical fits. The amplified seasonal evolution of the error in Fig. 6 relative to the $\Delta C_X^{\text{mapping}}$ cases is thus a representation of how temporal variability in given domains is captured differently by two different observational networks.

6.1.2 Constant formula strategy

The relative and absolute errors in the determination of the change in North Atlantic carbon inventory resulting from the use of eMLR with fixed regression structures (strategy 2) for both the GLODAP and CLIVAR datasets, for the hybrid and fixed network cases, projected either backward or forward in time onto the corresponding stations, are shown on Fig. 7 for all 255 first-order models. Eighty-five percent of all models tested yield basin-integrated $\Delta C_X^{\text{hybrid}}$ estimates that are within 20 % of the true value, with 73 and 76 % resulting in an underestimation when results are mapped from the GLODAP or CLIVAR stations. For the $\Delta C_{\text{GLODAP}}^{\text{fixed}}$ and $\Delta C_{\text{CLIVAR}}^{\text{fixed}}$ cases, these values change to 99 and 97 % below the 20 % error and 64 and 80 % resulting in underestimation. The mean relative errors across all models are -7.5 % (for $\Delta C_{\text{GLODAP}}^{\text{hybrid}}$), -5.3 % (for $\Delta C_{\text{CLIVAR}}^{\text{hybrid}}$), -2.0 % (for $\Delta C_{\text{GLODAP}}^{\text{fixed}}$) and -4.3 % (for $\Delta C_{\text{CLIVAR}}^{\text{fixed}}$). All means are significantly different from 0 (two-tailed t test, $p < 0.001$). Estimates obtained from projecting the hybrid results either backward or forward in time on the GLODAP or CLIVAR stations are well-correlated (Pearson' $\rho = 0.93$, $p < 0.001$), confirming that the influence of mapping errors is small when considering basin-scale inventories. The correlation is slightly less strong for the fixed-network cases (Pearson' $\rho = 0.85$, $p < 0.001$).

The across-model average underestimations of the 10 yr inventory change is -0.3, -0.22, -0.08 and -0.18 Pg C in absolute terms for the GLODAP and CLIVAR hybrid-network cases and the GLODAP and CLIVAR fixed-network cases. Most results easily meet the LSCOP criterion (Bender et al., 2002) presented in the introduction for the North Atlantic as 51 %, 60 %, 87 % and 89 % percent of the regression formulae yield results within 0.5 Pg C of the true estimate. Considering that about one third of the global carbon inventory is in the North Atlantic (Steinfeldt et al., 2009) and assuming a global 3 Pg C increase over 10 yr, 1 Pg C is proposed as a North Atlantic target over 10 yr: 91 %, 93 %, 99 % and 98 % percent of the models tested produce North Atlantic carbon inventory change estimates within 1 Pg C of the true value. It is clear that most regression models produce estimates of the integrated basin-scale decadal inventory change that meet desired accuracy limits. Results are consistently better when the observational network is fixed in time than

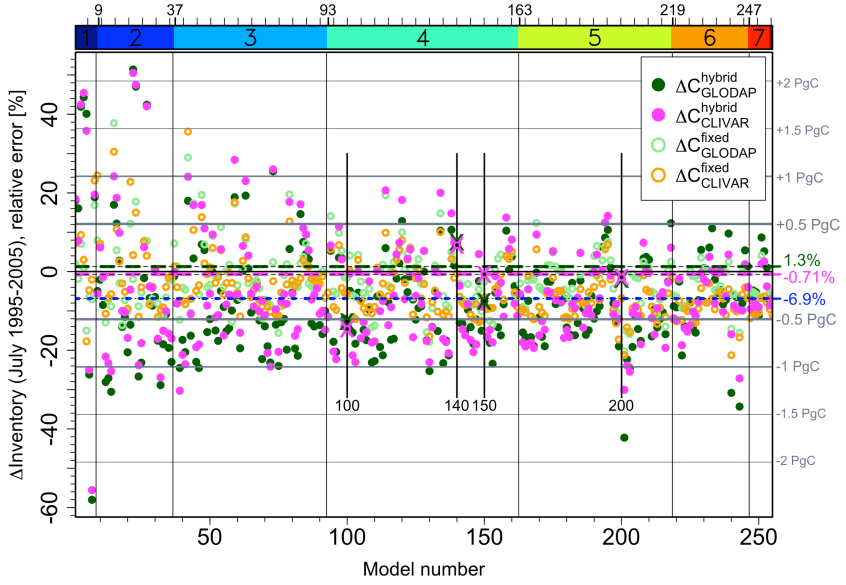

Fig. 7. Relative (left y-axis) and absolute (right y-axis) error in the change in anthropogenic carbon inventory for strategy 2 (constant model structure for all layers) between July 1995 and July 2005 for all possible 255 first order linear models. Hybrid results obtained by using combinations of regression models specific for 1995 GLODAP and 2005 CLIVAR sampling networks and projected either onto the 1995 GLODAP or the 2005 CLIVAR data ($\Delta C_X^{\text{hybrid}}$) are shown in dark green and magenta. Parallel results obtained by using combinations of models specific for each sampling networks when holding the observational networks fixed in time and projected onto the 1995 GLODAP or the 2005 CLIVAR data ($\Delta C_X^{\text{fixed}}$) are shown in light green and orange. Model size is indicated by the color strip on top. Mapping errors calculated from the "true" values are show as the horizontal dashed green (1.3 %, GLODAP) and magenta (−0.71 %, CLIVAR) lines. The dotted horizontal blue line (−6.9 %) shows the relative error when calculating the inventory change between 2005 and 1995 using the contemporary carbon fields, without removing natural variability.

when it varies, though. The fixed-GLODAP cases produce the best results overall.

All models with 7 or more terms and all the composite best-AIC solutions (strategy 1) for every month (Fig. 6) produce estimates that are better than the 0.5 Pg C error limit and thus exceed the success criterion proposed in the LSCOP report (Bender et al., 2002). Simpler models, such as models Z^{140} and Z^{150}, which stood out particularly in Figs. 2 and 3, also fall within the 0.5 Pg C accuracy limit. In fact, looking at the formula structure of the best 20 models with respect to how close their predicted inventory changes are to the true value indicates that, across all fixed-network and hybrid-network cases, 5 to 7 4-term models are present in this list consistently, along with 1 to 7 3-term models and 3 to 6 5-term models. Interestingly, there is only one 7-term model in the top-20 list and no 8-term or 1-term model, indicating that some intermediate complexity regression models can apparently outperform more complex models in estimating the basin-integrated carbon change.

Comparing the change in carbon inventory implied by the contemporary carbon field (Fig. 1e) without any effort to correct for natural variability (Fig. 1f) to the true anthropogenic carbon change (Fig. 1a) would result in only a −6.9 % error. This would also be within the acceptable limits. In fact, this result is better than 68 % of either GLODAP and CLIVAR hybrid-network eMLR results and better than 33 % and 51 % of the GLODAP and CLIVAR fixed-network solutions. Although regression-model selection does not appear to be a fundamental concern when the goal is to calculate the decadal basin-scale inventory change, comparing Fig. 1a and f clearly indicates that a small 6.9 % error actually corresponds to relatively large differences in the distribution of the recovered carbon change signal: it represent the error pattern shown in Fig. 1f. Since positive and negative errors cancel each other, basin-integrated measures are not particularly sensitive tests of quality. The next sections contrast results obtained by different regression models and show they can also result in substantially different interior distribution even if their integrated inventory change estimates are close to the true value.

6.2 Layer-specific inventories

Vertical profiles of the absolute errors of the 1995 to 2005 layer-specific inventory changes calculated by eMLR layer-by-layer for all first-order models (strategy 2) for the $\Delta C_{\text{GLODAP}}^{\text{hybrid}}$ case are shown in Fig. 8a. Although the magnitude of the errors vary slightly when the results are mapped from eMLR results projected on CLIVAR stations, the shape of the layer-specific inventory change profiles are mostly similar between CLIVAR and GLODAP results.

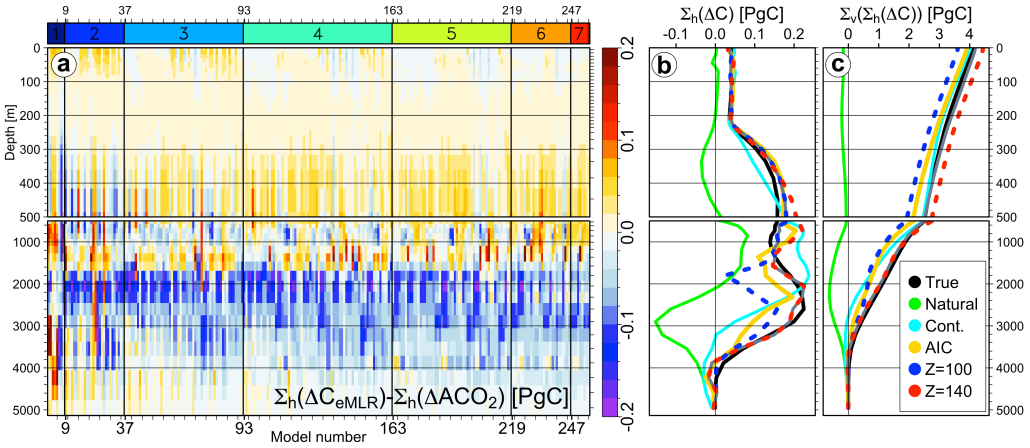

Fig. 8. (a) Absolute errors between the North Atlantic eMLR predicted inventory change, mapped from estimates at GLODAP stations, and the true inventory changes integrated on each horizontal model layer (Σ_h) and for each first order regression model (strategy 2). **(b)** Vertical profiles of the layer inventory changes and **(c)** vertically integrated layer inventory change (from the bottom to the surface, Σ_v). The true, natural and contemporary (Cont.) layer inventory changes between July 1995 and July 2005 are shown, together with the "best AIC" composite solution and results from models Z^{100} and Z^{140} (dotted) and their merged products spliced at 1500 m (gray).

The general shape of the fixed-network layer-specific inventory error profiles also tends to be similar to the hybrid-network results, as indicated by the dominantly linear relationships in Fig. 9a and b. Deviations from the 1 : 1 line in Fig. 9a and b show that the large errors tend to be generally smaller in the fixed-network cases than in the hybrid cases, however. There is no relationship between the differences in layer specific inventories between the hybrid and fixed network cases and the absolute errors of the fixed-network results (Fig. 9c, d). An inverse linear relationship exists, though, between the fixed-hybrid differences and the absolute error calculated between the hybrid results and the true values (Fig. 9e ,f). This relationship is mostly due to points in the depth range 300–2300 m as indicated by the color-code of the points in Fig. 9e and f. The discontinuity in the 600–1000 m layer visible in the hybrid results (Fig. 8a) does not exist in the $\Delta C^{\text{fixed}}_{\text{GLODAP}}$ results. $\Delta C^{\text{fixed}}_{\text{GLODAP}}$ estimates tend to consistently overestimate, not underestimate, the true values in that zone, as is the case above and below that layer. This is less clear in the $\Delta C^{\text{fixed}}_{\text{CLIVAR}}$ case. Although the $\Delta C^{\text{fixed}}_{\text{GLODAP}}$ and $\Delta C^{\text{fixed}}_{\text{CLIVAR}}$ results are well correlated, most discrepencies between these two cases are in the depth range 500–1700 m, indicating that network differences have the most influence on the results on these horizons.

Aspects of the vertical profiles in Fig. 8 are consistent with features of the AIC profiles in Fig. 5. One notable similarity is the band of relative AIC highs centered around 2000 m (Fig. 5a, c) which coincides with a layer of systematically strong underestimation (Fig. 8a). The region between 1500 to 3500 m is where most of the error (underestimation, Fig. 7) in the basin-scale inventory change estimates is generated (Fig. 8c). This is true for all fixed-network and hybrid-network cases.

The 1500–3500 m layer also displays large changes in the natural carbon component (green line, Fig. 8b and c). Closer inspection of the model tracer fields indicates this is primarily a reflection of variations in convective activity and associated water mass reorganization. The changes in model tracers consistently point to water mass aging in the deeper layers due to shallower convective mixing in 2005 relative to 1995, when deeper waters were better ventilated.

Similar to the basin-integrated inventory changes (Fig. 7), the $\Delta C^{\text{AIC, hybrid}}_{\text{GLODAP}}$ case (yellow line, Fig. 8b and c) does not produce the most accurate profile. Figure 8b and c show that an eMLR solution spliced from simpler 4-term models, namely those identified in Fig. 2 and displaying a high degree of vertical continuity (Z^{100} in the upper 1500 m and Z^{140} below that depth), can reproduce the true layer-specific inventory change profile almost exactly. As indicated in Fig. 2, the family of 4-term models shows strong vertical continuity between layers, with essentially four formulae able to cover all depths from 100 to 4000 m. Specifically, these 4-term models (numbers 140, 99, 100 and 150, ordered by the relative frequencies with which they are selected) contain the following predictor variables: $Z^{140} = \{\theta, PO_4, Si, Alk\}$, $Z^{99} = \{S, \theta, PO_4, AOU\}$, $Z^{100} = \{S, \theta, PO_4, O_2\}$, $Z^{150} = \{NO_3, PO_4, Si, Alk\}$.

Model Z^{140} is the model structure used by Friis et al. (2005) for their North Atlantic analysis and Z^{100} is the model used by Levine et al. (2008). Interestingly, while Levine et al. (2008) applied this formula globally to model fields between 200 and 2000 m on every grid point, Fig. 2 suggest this formula is more appropriate in the upper 200 m given the GLODAP station coverage. Models Z^{99} and Z^{100} are nearly identical, the only difference between the two being the use of O_2 or AOU. Model Z^{150}, which fits the data

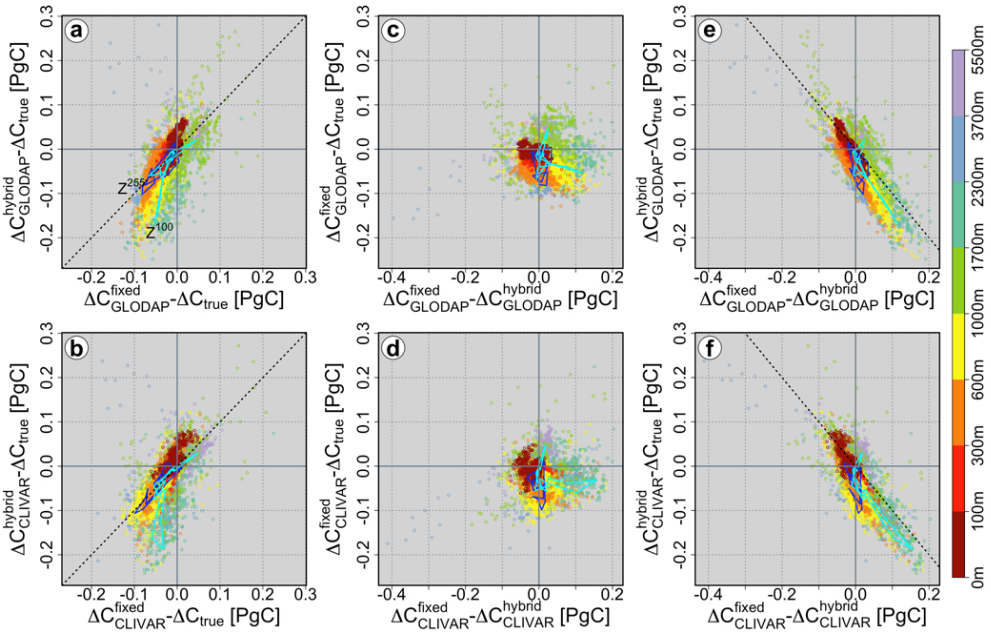

Fig. 9. Relationship between the absolute error in the layer-specific inventories obtained by the fixed-network or hybrid network cases for the **(a)** GLODAP and **(b)** CLIVAR stations. Relationships between the fixed-hybrid differences in layer-specific inventory change estimates and the absolute error between the fixed and true estimates for **(c)** GLODAP and **(d)** CLIVAR. Relationships between the fixed-hybrid differences in layer-specific inventory change estimates and the absolute error between the hybrid and true estimates for **(e)** GLODAP and **(f)** CLIVAR. The color scale identifies points from depth interval representative of the layering in Fig. 8a.

well in the range 2000 to 4000 m is interesting in that it does not include either θ or S in its formula. This reflects the fact that the dynamic range of these tracers is small in that depth range relative to that of other tracers (Appendix A). This is qualitatively consistent with the classic studies of Broecker (1974) and Broecker et al. (1985) who relied on nutrient-based composite quasi-conservative tracers ("NO", "PO") to characterize the flow path of deep waters in the Atlantic.

While some of the models identified from the GLODAP analysis (Fig. 2) are also present in the CLIVAR analysis (Fig. 3), their vertical stacking can differ. This is the case for models Z^{99}, Z^{100} and Z^{140}. Given the CLIVAR stations, Z^{140}, the model of Friis et al. (2005) takes a prominent role in the top 200 m while models Z^{99} and Z^{100}, the model of Levine et al. (2008), occupy the space between 300 and 500 m. Models $Z^{97} = \{S, \theta, NO_3, Alk\}$ and $Z^{98} = \{S, \theta, PO_4, Si\}$ belong essentially to the same model group as Z^{99} and Z^{100} as all these models feature salinity, temperature and phosphate (or nitrate) as dominant variables. Z^{97}, Z^{98} extend the influence of this model group down to about 3000 m, although the continuity is not as clear as with models Z^{140} or Z^{150} in the GLODAP case.

The differences between Friis et al. (2005) and Levine et al. (2008) can be explained by the results of our analysis. Friis et al. (2005) performed their analyses on data located in the Subpolar North Atlantic (North of 40° N, South of Iceland) and many of the data used in Friis et al. (2005) are the same data that partly constitute GLODAP (Key et al., 2004) in that region. It is then reassuring that both Friis et al. (2005) and our results converge towards the same model (Z^{140}) for the appropriate size-class. Similarly, the synthetic model dataset used by Levine et al. (2008) was most heavily influenced by the subtropical regions. This is because Levine et al. (2008) included every grid box in their analysis and did not subsample their model to mimic the station coverage of the observational datasets. In that sense, the spatial bias of their dataset is more like CLIVAR, and it is again reassuring that model Z^{100}, or related models, be most representative in these two cases.

Given that this analysis uses a physical and biogeochemistry model as a source of data, that Levine et al. (2008) used a different circulation and biogeochemistry model and that Friis et al. (2005) used observations, it is encouraging to note how well the regression formulae proposed by each study converge when presented in the context of their sampling grids. Whether a simple combination of the regression formulae $Z^{100} = \{S, \theta, PO_4, O_2\}$ and $Z^{140} = \{\theta, PO_4, Si, Alk\}$, as indicated in Fig. 8b and supported by Fig. 2, is appropriate for application of eMLR to the real dataset remains to be seen. Based on the analysis of the layer inventories, the fact that the TOPAZ model is a state-of-the-art biogeochemistry model and the robust correspondence with other studies, it would appear, however, that these are a priori good candidate formulae in the North Atlantic.

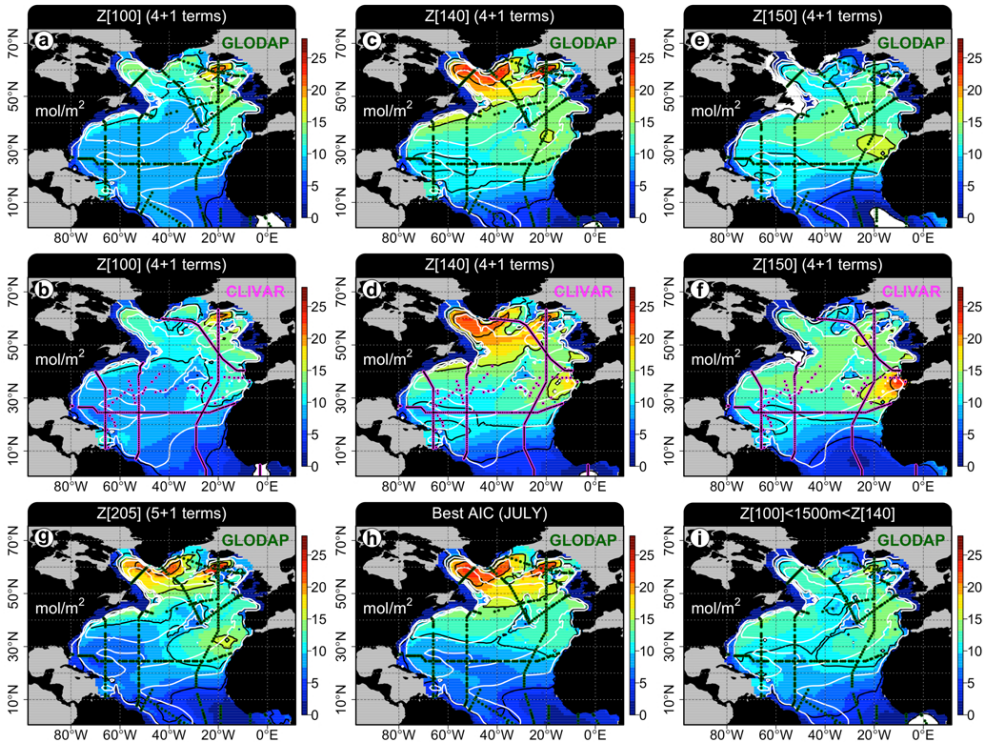

Fig. 10. Calculated eMLR anthropogenic carbon column inventory change between July 2005 and 1995 mapped from either the GLODAP (green) or CLIVAR (magenta) stations when the sampling networks change in time ($\Delta C_X^{\text{hybrid}}$). Black contours are drawn in increment of 5 mol m^{-2}. White contours reproduce the shape of the best possible pattern obtained from mapping the true values sampled at the station locations (1b). Results are shown for models **(a, b)** $Z^{100} = \{S, \theta, \text{PO}_4, \text{O}_2\}$, **(c, d)** $Z^{140} = \{\theta, \text{PO}_4, \text{Si, Alk}\}$, **(e, f)** $Z^{150} = \{NO_3, PO_4, Si, Alk\}$, **(g)** $Z^{205} = \{\theta, \text{NO}_3, \text{Si, AOU, Alk}\}$, **(h)** for the "best AIC" models selected by minimum AIC for each layer and each time point (strategy 1) and **(i)** the merged product using Z^{140} below 1500 m and Z^{100} above.

Obviously, absolute errors cannot be used as guides for model selection when working with real data. Based on the analysis of the layer inventories, the criterion of vertical continuity of statistically selected models can seemingly be used to guide model selection and define their extent of use vertically in conjunction with a general oceanographic assessment of the regression residuals. As shown in Fig. 8b, these criteria can be applied to isolate model formulae that perform as well or better than more complex formulae when evaluated at the basin-scale inventory level or when looking at layer-specific inventory change profiles. Layer inventories integrate over large horizontal scales, however, and in the direction (horizontal) along which the regression fitting is performed. Because regression analysis is designed to minimize the distance of the data relative to the mean, if the data are symetrically distributed around the layer mean value, it is likely that positive and negative residuals cover more or less equal areas and largely cancel upon integration. Layer-inventories can then underestimate potential problems. The next section contrasts horizontal inventory changes with inventory changes calculated vertically, perpendicular to the direction along which the regressions are performed.

6.3 Column inventories

Figure 7 shows that acceptable basin-integrated inventory change estimates can be obtained from different regression formulae. Yet, these formulae produce vertical profiles of the layer-specific anthropogenic carbon inventory change that can vary and yield biases in the ocean interior (Fig. 8). This section presents a complementary view, investigating the geographical distribution of the eMLR-calculated column inventory change, the associated absolute error patterns and their correlations with the vertically integrated true signal and the natural variability pattern.

Illustrative column inventory change estimates for the $\Delta C_X^{\text{hybrid}}$ case resulting from the use of fixed model structures at all depths but with regression coefficients optimized layer-by-layer and for the GLODAP and CLIVAR datasets independently (strategy 2) are shown in Fig. 10a–g. The corresponding absolute error maps calculated between the $\Delta C_X^{\text{hybrid}}$ estimates and the true change in column inventory (Fig. 1a) are shown in Fig. 11a–g. Results for the 4-term models $Z^{100} = \{S, \theta, \text{PO}_4, \text{O}_2\}$ (a, b), $Z^{140} = \{\theta, \text{PO}_4, \text{Si, Alk}\}$ (c, d) and $Z^{150} = \{\text{NO}_3, \text{PO}_4, \text{Si, Alk}\}$ (e, f), selected based on

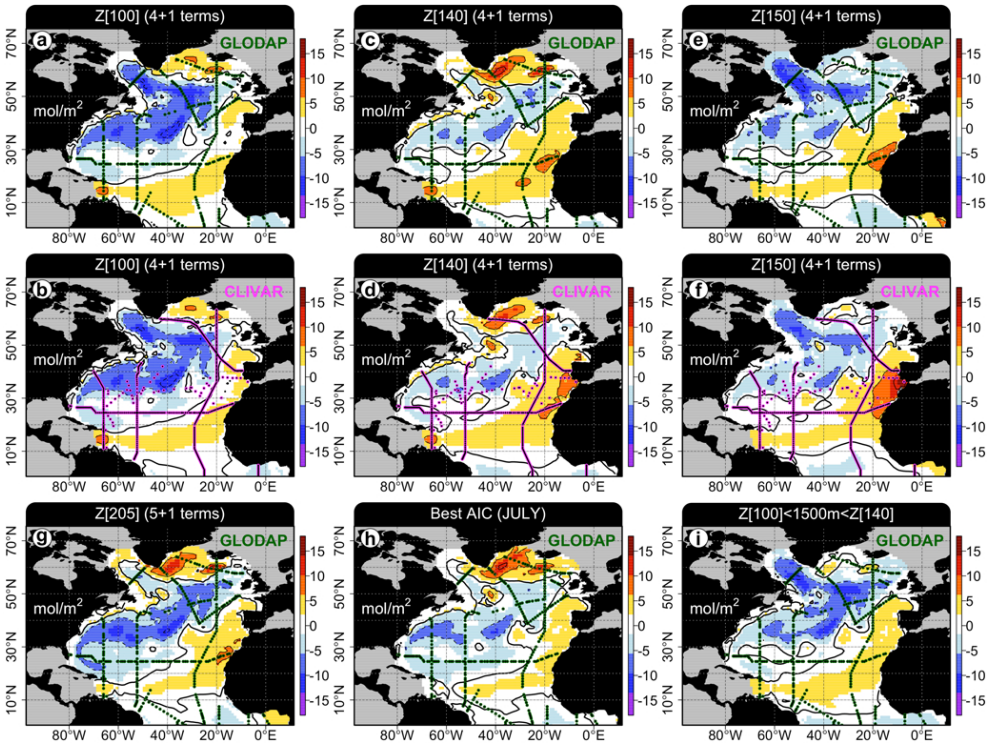

Fig. 11. Absolute errors in anthropogenic carbon column inventory changes between July 2005 and 1995 mapped from either the GLODAP (green) or CLIVAR (magenta) stations calculated from sampling networks that change in time ($\Delta C_X^{\text{hybrid}}$). Dashed (negative) and solid (positive) contours are drawn in increment of 6 mol m^{-2}. The thick lines marks the 0 contour. Results shown for models **(a, b)** Z^{100}, **(c, d)** Z^{140}, **(e, f)** Z^{150}, **(g)** Z^{205}, **(h)** for the "best AIC" models selected by minimum AIC for each layer and each time point (strategy 1) and **(i)** the merged product using Z^{140} below 1500 m and Z^{100} above.

Figs. 2 and 3, are shown on both figures. Parallel results obtained by holding the observational networks fixed in time ($\Delta C_X^{\text{fixed}}$) are shown in Figs. 12 and 13.

Differences in the vertically integrated patterns between $\Delta C_{\text{GLODAP}}^{\text{hybrid}}$ (Figs. 10 and 11a, c, e) and $\Delta C_{\text{CLIVAR}}^{\text{hybrid}}$ (Figs. 10 and 11b, d, f) are small relative to differences in error patterns observed between results generated from different regression formulae. This is also true in the fixed-network cases, but differences amongst the column inventories obtained by applying various regression models in the fixed-network cases (Fig. 12) are much smaller than differences amongst estimates for the hybrid cases. These examples show that regression model selection has a much greater influence on the final results when the location of the stations change in time than when the observational network is constant. This is a consequence of the systematic misfit cancelation effect.

The root-mean-square error (RMSE) of column inventories obtained by mapping the true results sampled at either GLODAP or CLIVAR stations (Fig. 1c, d) can be thought of as the best realizable RMSE given the stations available. The RMSE due to mapping only is about half (≈ 1.5 mol m^{-2}, Fig. 14) the RMSE of the best eMLR hybrid-network results (≈ 3 mol m^{-2}). The smallest fixed-network RMSE (≈ 2 mol m^{-2}) are closer to the RMSE due to mapping only (Fig. 15). In comparison, the RMSE of the vertically integrated natural variability pattern in Fig. 1f is 4.2 mol m^{-2}. This means that even if not all the natural variability is accounted for by eMLR (there remains an offset relative to the mapping-only results), hybrid-network results account for about 45 % of it, while fixed-network solutions can remove about 81 %.

The magnitude of RMSE for hybrid and fixed-network results and the range of RMSE values across all models (4–5 mol m^{-2}) confirms that, in the North Atlantic and when the observational networks changes in time, mapping is not the dominant factor controlling the basin-scale structure of the error maps shown in Fig. 11, even if mapping errors can be locally significant in unsampled dynamic regions. The influence of mapping is relatively more important when the observational network is fixed in time, however. In that case, because the results are more accurate, a large fraction (about 75 %) of the total RMSE is due to the "bull's eyes" in Fig. 12, some of which are due to mapping and undersampling (Figs. 1c, d). Calculating the errors relative to the mapped true values (Fig. 1b) effectively removes the "bull's eyes" off the Grand Banks and in the North American Basin. This does not address the overestimate in the Irminger Sea

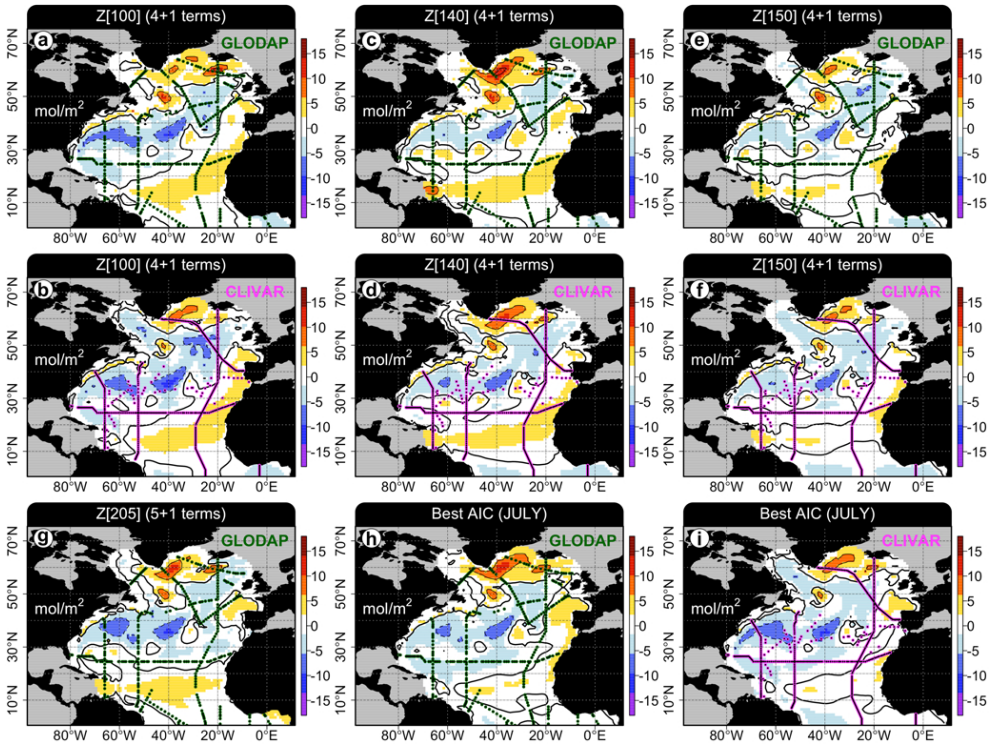

Fig. 12. Absolute errors in anthropogenic carbon column inventory changes between July 2005 and 1995 mapped from either the GLODAP (green) or CLIVAR (magenta) stations calculated from sampling networks that do not change in time ($\Delta C_X^{\text{fixed}}$). Thin dashed (negative) and solid (positive) contours are drawn in increment of 6 mol m^{-2}. The thick lines marks the 0 contour. Results shown for models **(a, b)** Z^{100}, **(c, d)** Z^{140}, e, f: Z^{150}, **(g)** Z^{205}, and for cases when the "best AIC" models are selected for each layer and each time point (strategy 1) for **(h)** GLODAP and **(i)** CLIVAR.

and does not alter the main structure of the error maps, however.

Results from the 8-term model Z^{255} tend to be similar to the "best AIC" composite estimates from strategy 1 (Figs. 10h, 11h and 12h, i). This is because models with the lowest overall AIC values also tend to be the more complex ones in the absence of analytical uncertainties (Figs. 2 and 3). This can be seen in Fig. 14a, b and d where the points corresponding to Z^{255} systematically overlap the AIC cluster. The tight clustering of the AIC results in Fig. 14 is due to the small influence that interannual changes in seasonality has on the results relative to the errors induced by regression.

The regression model producing the smallest overall RMSE for $\Delta C_{\text{GLODAP}}^{\text{hybrid}}$ and $\Delta C_{\text{CLIVAR}}^{\text{hybrid}}$ (Fig. 14) is 5-term model $Z^{200} = \{\theta, \text{NO}_3, \text{PO}_4, \text{Si}, \text{Alk}\}$. The relative success of model Z^{200} in the hybrid-network case would be hard to predict based uniquely on the fit statistics. Figures 2 and 3 show Z^{200} is only selected as a best-fit model in its size-class on a few deep horizontal layers and in the GLODAP case only. Fit Z^{200}, in spite of not yielding the smallest residuals, is consistently very good as large portions of the water column have AIC values within 10 % of the minimum AIC for this model (Fig. 5). Column inventory changes from model Z^{200} (not shown) are similar in magnitude and structure to results from model Z^{140} and to the "best-AIC" case. Differences between these results are small and regional.

Although model Z^{200} is closely related to model $Z^{205} = \{\theta, \text{NO}_3, \text{Si}, \text{AOU}, \text{Alk}\}$ (Fig. 10g) that was used in the observational study of Tanhua et al. (2007), who analyzed zonally oriented data in the subtropical North Atlantic, the magnitude of the column inventories predicted by these two models differ strongly when applied over the basin-scale. The replacement of PO_4 in Z^{200} with AOU in Z^{205} results in much larger underestimations in the whole zone between the Caribbean and Ireland (Fig. 11g). Formula Z^{205} is never selected as a best-fit model given the station coverage considered here (Figs. 2 and 3). Z^{200} also results is better column inventory estimates than Z^{205} when the observational network is fixed (Fig. 12g), but differences are small.

Model Z^{100}, which in contrast to Z^{140} does not use Si or Alk but rather S and O_2 in addition to PO_4 and θ as predictor variables, produces smaller overestimates than model Z^{140} in the subequatorial region, the eastern Atlantic and in the Irminger Sea in the hybrid-network cases but Z^{100} results in inflated and extended underestimations over the North American Basin and the Labrador Sea (Figs. 10a, b and 11a, b), producing greater overall RMSE (Fig. 14). Differences

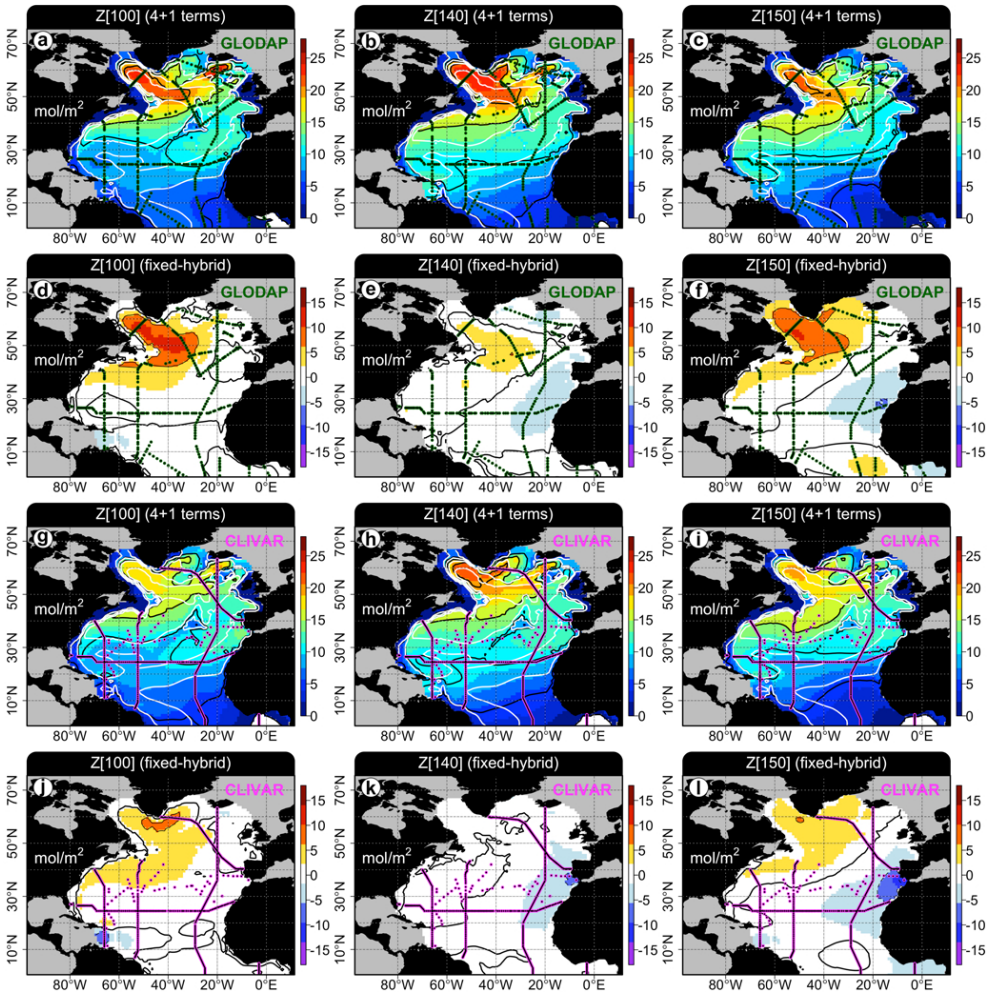

Fig. 13. Mapped anthropogenic carbon column inventory changes calculated from sampling networks that do not change in time ($\Delta C_X^{\text{fixed}}$): **(a, g)** Z^{100}, **(b, h)** Z^{140} and **(c, i)** Z^{150}. In these panels, black contours are drawn every 5 mol m^{-2}; white contours reproduce the shape in Fig. 1b. Differences between the fixed and hybrid cases ($\Delta C_X^{\text{fixed}} - \Delta C_X^{\text{hybrid}}$) are shown for GLODAP **(d, e, f)** and CLIVAR **(j, k, l)**; contours drawn every 6 mol m^{-2} with a thick line at 0 mol m^{-2}.

between Z^{100} and Z^{140} are smaller in the fixed-network case (Figs. 12a–d and 13a–b, g–h).

A composite result built from Z^{100} in the top 1500 m and Z^{140} below that depth (Fig. 10i), as suggested earlier based on the analysis of the layer inventory change profile (Fig. 8), produces a column inventory change map that is yet different from all other examples shown in Fig. 10. Although this solution improves the signal along the western boundary relative to the Z^{100} pattern (Fig. 10a) and results in an accurate quantification of the layer-specific inventory change profile (Fig. 8) and of the basin-scale inventory change, the large signal in the Labrador Sea and subpolar region is missing. The RMSE of this composite case is intermediate between that of Z^{100} and Z^{140} (asterisks, Fig. 14).

Nutrient-based model Z^{150} uses neither of the physical parameters θ, S nor O_2, and produces yet another inventory pattern in the hybrid-network case (Fig. 10e, f), with RMSE intermediate between Z^{140} and Z^{100} (Fig. 14). The error patterns for Z^{150} are typified by large-scale overestimations over the subequatorial and eastern Atlantic and large-scale underestimations over the Northwest Atlantic (Fig. 11e, f). This is in drastic contrast with the column inventory produced by Z^{150} in the fixed-network case (Fig. 12e, f and 13c, i). Z^{150} is the model that produces the smallest RMSE in that case (Fig. 15).

Although RMSE values between models Z^{100}, Z^{140} and Z^{150} vary by less than 1 mol m^{-2} and each represents the best-fit regression models over substantial parts of the water column (Figs. 2 and 3), the dynamical interpretation that would be associated with the eMLR column inventories

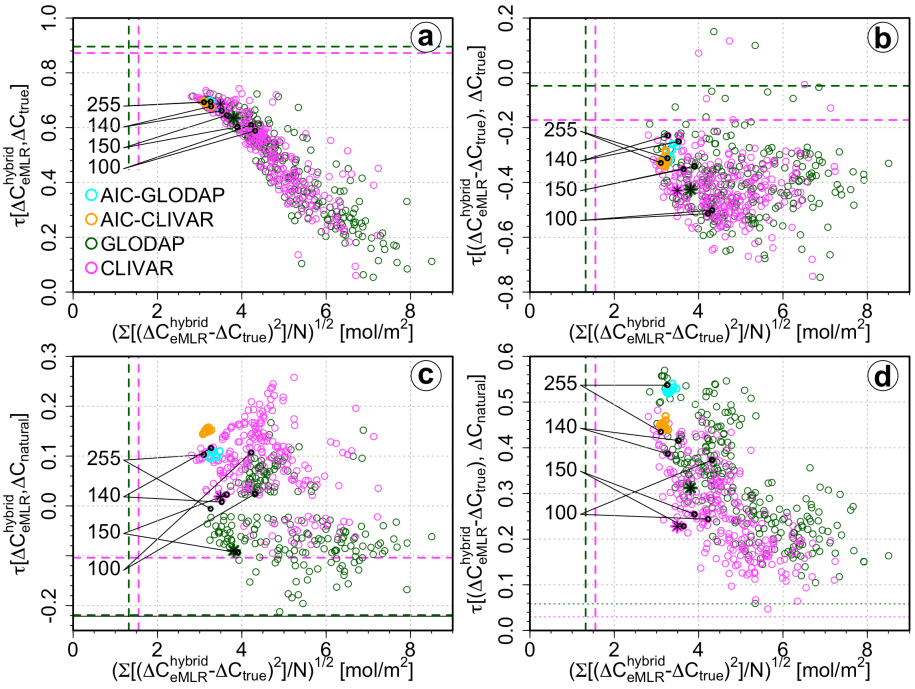

Fig. 14. Kendall correlation coefficients (τ) calculated between the column inventory changes predicted by all eMLR solutions obtained with a variable sampling network (strategy 1 and 2, $\Delta C_X^{\text{hybrid}}$) and **(a)** the true or **(c)** the mapped natural carbon change plotted as a function of each model's RMSE relative to the true values. **(b, d)** Corresponding correlations calculated using the absolute error patterns. Only correlations with $p \leq 0.05$ are plotted. The RMSE due to mapping only are shown by vertical green (GLODAP) and magenta (CLIVAR) lines. Correlations calculated using the true mapped values instead of the eMLR results are shown as horizontal green (GLODAP) and magenta (CLIVAR) lines; thin dotted lines in **(b)** mean correlations are not significant at $\alpha = 0.05$. The solid horizontal gray line in **(c)** shows the correlations between the true values (no mapping) and the natural carbon change pattern.

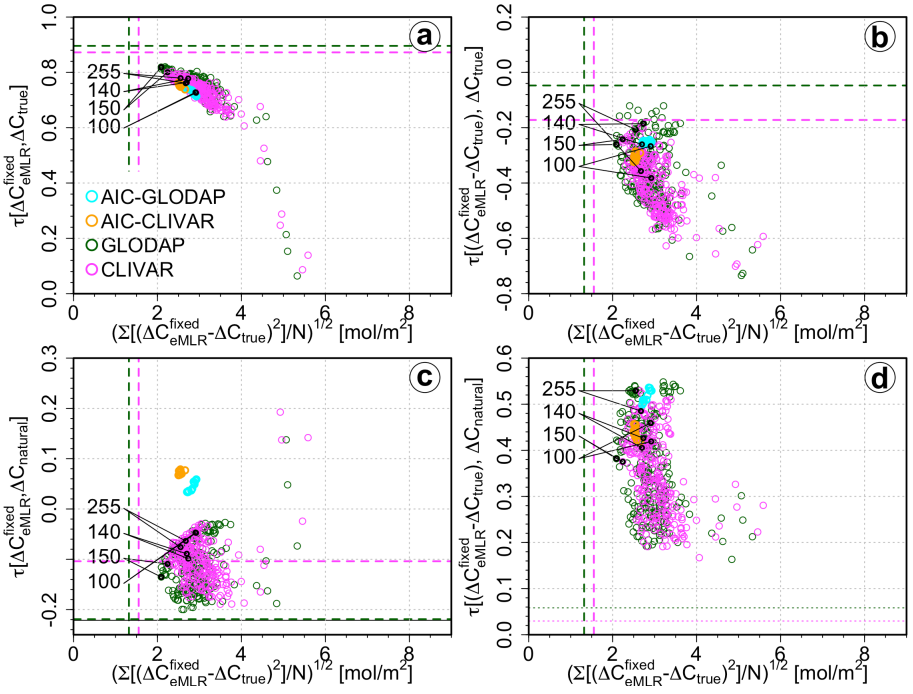

Fig. 15. Same as Fig. 14 when sampling networks are held fixed in time ($\Delta C_X^{\text{fixed}}$).

produced by these models (Fig. 10) when the observational network changes in time vary dramatically. Of these three examples, carbon changes produced by model Z^{140} are closest to the true signal (Fig. 1a), indicating greatest uptake in the Labrador Sea, elevated values that extend zonally eastward around 50° N, and southward propagation along the western boundary (Fig. 10c, d). In contrast, results from Z^{150} show weak basin-scale gradients with little subpolar-subtropical difference (Fig. 10e, f). These hybrid-network results also exaggerate the carbon change west of Gibraltar while grossly underestimating the role of the Labrador Sea. Results from model Z^{100} would even indicate a clockwise northeastern boundary intensification of the column inventory change (Fig. 10a, b) that is inconsistent with the vertically integrated changes of the true signal (Fig. 1a).

Results produced by the fixed-network cases are more consistent with each other and with the true carbon inventory change pattern than the hybrid-network results. The greatest differences between the hybrid and fixed-network results occur in the subpolar gyre region, where column inventory changes predicted by $\Delta C_X^{\text{fixed}}$ are typically larger than those predicted by the $\Delta C_X^{\text{hybrid}}$ cases (Fig. 13d–f, j–l). This is the manifestation of the mid-depth (500–1700 m) differences seen in the layer-specific inventory changes in Fig. 9e–f. In that region, the fixed-hybrid differences are smaller when mapped from CLIVAR stations (Fig. 13j–l) than mapped from the GLODAP stations (Fig. 13d–f). The fixed-network results also tend to produce smaller changes in column inventories than the hybrid cases in the Eastern Atlantic, between the coasts of Portugal and Senegal. This fixed-hybrid difference is larger in the CLIVAR case than in the GLODAP case (Fig. 13d–f, j–l).

Clearly, eMLR-derived column inventory change patterns depend on the choice of regression model used in the calculation, even if most results produce basin-integrated inventory change estimates within tolerated accuracy limits. The variability between results and the influence of regression model selection is strongly modulated by the temporal evolution of the observational network. Solutions obtained when the observational network is fixed in time are better and less variables than when the spatial sampling density varies. None of the regressions tested yield totally unbiased results but eMLR results appears to converge towards a particular large-scale error pattern. This pattern is characterized by an overestimation of the column inventory in the Irminger Sea and in the Eastern subtropical Atlantic, and an underestimation in the Western subtropical Atlantic and along the path of the North Atlantic Current. Both layer-inventories and column inventories point towards the intermediate and deep subpolar region as the root cause for the biases. While some formulae produce column inventory patterns that are mostly similar to the true signal and that can be readily interpreted, others result in patterns whose interpretation would lead to significant misunderstanding of the penetration pathway of anthropogenic carbon.

6.4 Correlations

The degree of pattern similarity between the mapped eMLR column inventory changes, the corresponding error patterns, the true signal and the natural variability component are further examined here using non-parametric Kendall correlation analysis. Correlation coefficients (τ) with p-values smaller than 0.05 are plotted in Figs. 14 and 15 as a function of the RMSE of column inventory change estimates.

6.4.1 Correlations between eMLR estimates and the true anthropogenic signal

As expected, eMLR column inventory changes with lower RMSE are better correlated with the true simulated carbon change map (Figs. 14a and 15a). Correlation coefficients tend to plateau around $\tau \approx 0.7$ when RMSE is smaller than 3.7 mol m^{-2} in the hybrid-network case, however. Maximum correlations are slightly higher, around $\tau \approx 0.8$, in fixed-network cases. This plateau effect reflects convergence of the eMLR solutions towards a particular pattern, i.e. the broad-scale pattern captured in Figs. 10c, d, g, h and 13a–c, g–i. Changes in RMSE below ≈ 3.7 mol m^{-2} increasingly represent regional scale differences, while results beyond that breakpoint in slope (Figs. 14a and 15a) largely represent major differences in basin-scale patterns.

Ideally, no correlation should exist between the absolute error and the true signal since the error should be zero everywhere. Nonetheless, a weak negative correlation is introduced between the error and the true signal (Fig. 14b) by the mapping process and the fact that high-change regions are undersampled and thus underestimated. There is no clear relationship between the correlation coefficients and RMSE for the hybrid-network case (Fig. 14b), although the variance amongst τ-values tends to be less for cases with lower RMSE. Correlation results from the fixed-network cases show less scatter than their hybrid counterparts and do point toward a systematic reduction of the correlation between the error with the true signal with decreasing RMSE (Fig. 15b). The dominantly negative correlations in Fig. 14b and 15b are of greater magnitude than the mapping-only τ-values because most eMLR solutions tend to produce overestimates in large regions of the North Atlantic, principally at low latitudes and in the Eastern half of the basin, regions characterized by a small true column inventory change signal (Fig. 1a) while also underestimating the carbon change over the North American Basin and in parts of the subpolar region (Fig. 11), two regions characterized by large carbon changes (Fig. 1a).

6.4.2 Correlations between eMLR estimates and the change in natural carbon

Although being controlled by different processes and having different large-scale patterns (Fig. 1a, f), a weakly negative correlation exists between the true column inventory change pattern and the pattern of change in natural carbon (Fig. 14c). This weak correlation is entirely due to an area over the North American Basin where a large negative natural carbon change occurs (Fig. 1f) in the same region where a moderately large anthropogenic carbon increase is seen. Correlations between the eMLR-estimated column inventory changes and the natural component are similarly weak in magnitude and many are not statistically significant (τ values for GLODAP versions of Z^{255} and Z^{140} are not significant but are drawn for completeness in Fig. 14c).

Correlations in Fig. 14c do not vary systematically with RMSE, but there is a trend towards convergence of the correlation coefficient at lower RMSE ($< 3.7\,\mathrm{mol\,m^{-2}}$). Correlations calculated from maps created either from the GLODAP or CLIVAR stations are clearly offset in the hybrid-network case, with CLIVAR maps resulting mostly in weakly positive correlations. The magnitude of this offset is similar to the spread in correlations between eMLR results derived from the same observational networks. Much of the spread in Fig. 14c originates over the North American Basin (large positive change in anthropogenic carbon, large negative change in natural carbon) and the Irminger Sea (small change in anthropogenic carbon, large positive change in natural carbon). There is no GLODAP-CLIVAR offset between the correlation coefficients for the fixed-network cases (Fig. 15c) and these results are of similar magnitude as the correlations induced by mapping only. Results for the "best-AIC" strategy are decoupled from Z^{255} in Figs. 14c and 15c, in spite of these results producing closely related column inventory patterns. This decoupling is further evidence that the correlations in Figs. 14c and 15c are mostly driven by small overlap regions.

Just as there should be no correlation between the absolute error maps and the true change in anthropogenic carbon, there should be no correlation between the error maps and the vertically integrated change in natural carbon. No significant correlation exists between the mapping errors (e.g. the best possible eMLR solutions, Fig. 1c, d) and the natural carbon pattern (Fig. 1f), as expected (dotted lines are used in Fig. 14 to show lack of significance). Yet, as suggested by a visual comparison of the error patterns in Figs. 11 and 12, the degree of correlation between the error maps of the eMLR solutions and the vertically integrated natural carbon change is systematically positive and significant (Figs. 14d and 15d). The correlation coefficients vary substantially between regression models and appear to be inversely proportional to RMSE in the hybrid-network case (Fig. 14d). This relationships breaks down with decreasing RMSE, however (Figs. 14d and 15d); correlation coefficients can vary from about 0.2 to 0.55 for RMSE $< 3.7\,\mathrm{mol\,m^{-2}}$.

The significance of the correlations in Figs. 14d and 15d indicates that the large-scale natural variability pattern is never fully corrected for by horizontal basin-scale eMLR, even when the most complex and statistically best-fitting models are used systematically (see AIC results in Figs. 14d and 15d). Correlation analysis performed layer-by-layer using the station predictions instead of the mapped results further confirm this point as the vertical profiles of τ-values are positive and significant at most depths and for every regression model. This systematic shortcoming of eMLR reflects the influence of secular trends on the tracer fields. Secular trends modify the large-scale property gradients in ways that are not reflected by the existing tracer distribution and consequently cannot be statistically represented by the empirical models derived from the spatial regressions. The variance of the correlations coefficients in Figs. 14d and 15d is a reflection of the effect of regression model selection.

7 Discussion

7.1 Unresolved temporal variability

The analysis performed here relied on snapshots of the ocean state taken either in 1995 or 2005, a situation which is overly idealistic as hydrographic sampling programs are never instantaneous. July was chosen to approximate the summer bias that exists in the real datasets, and the years 1995 and 2005 were selected as they represent peaks in real sampling intensity. This section discusses the possible effects of unresolved variability on the eMLR results.

7.1.1 Seasonality

These analyses have shown that the ability of regression models to fit the DIC data varies through the seasonal cycle. The summer to winter contrast in the standard error of the regressions for either the GLODAP or CLIVAR sampling is about $5\,\mu\mathrm{mol\,kg^{-1}}$. This effect is restricted to the upper water column, however. The GLODAP 1995 to CLIVAR 2005 differences in the standard errors of the residuals are typically smaller than 2 μmol/kg, but are mostly caused by differences in the sampling grid and not temporal changes, as comparisons with corresponding GLODAP 2005 and CLIVAR 1995 cases show. Even if the upper ocean contains large anthropogenic carbon concentrations, the volume is relatively small and changes in seasonality only result in $\approx 4\,\%$ fluctuations (Fig. 6) on the relative error of the basin-scale estimate of the inventory change. This seasonally varying error is small relative to the effect of regression model selection.

The seasonal inventory change estimates in Fig. 6 were derived using month-by-month comparisons, where January 1995 is directly compared to January 2005, etc. In practice, real datasets are composed of samples taken from different

seasons. The seasonal bias inherent in the data is not expected to change greatly between sampling campaigns, however. Unless the seasonal biases in sample distribution contained in real datasets were to change drastically (e.g. all winter versus all summer values), the seasonal sampling bias is unlikely to become a dominant source of error at the basin-scale. Additionally, given the available sample distribution, differences in representativeness of the sampling grids have a larger effect on the dataset variance (Fig. A2) and model selection than seasonal changes. Lastly, since regression misfits are largest in the summer and early fall and hybrid-network eMLR solutions produce lower errors in winter (Fig. 6), addition of winter data should result in an overall improvement of the fit quality, a consequence of reduced biogeochemical gradients during the winter and spring seasons due to more intense mixing, and in eMLR estimates. While seasonal effects can produce local extrema in residuals at particular near-surface stations, seasonal variability tends to be filtered out and is not expected to bias the change in carbon inventory estimates obtained by eMLR on the space and time-scales relevant to the Repeat Hydrography program.

7.1.2 Sub-monthly to inter-annual variability

The synthetic datasets were generated from monthly mean fields such that sub-monthly variability is filtered out by design. The magnitude of seasonal variability outweighs sub-monthly variability. Since seasonal variability is unlikely to introduce large errors in the decadal eMLR inventory change estimates, and since eMLR is a statistical method that relies on a large number of data points, sub-monthly variability is not expected to play a role as long as spatial covariances typical of these temporal scales are small relative to the domain size.

There exists an implicit relationship between the spatial scales of the system under study and the temporal scales that are smoothed out by regression. As such, sub-monthly perturbations would have to affect either an extensive coherent region or be extremely large to have noticeable effects on the regression statistics and the eMLR results. Considering data on the basin-scale for the regression analysis is then equivalent to filtering out temporal variability that is uncharacteristic of that scale and is averaged out. As WOCE and CLIVAR are separated by approximately a decade, consideration of large domains consistent with spatial patterns of interannual variability limits aliasing of shorter term variability.

Obviously, the same considerations place limits on the type of variability eMLR can be expected to remove. Interannual modes of variability that have spatial scales similar to the basin-scale can look like secular trends from the point of view of eMLR depending on the relative phasing of the perturbation when observed. For these modes, even if the observational network were to remain fixed in time, the systematic misfit cancelation effect may not be able to correct the inconsistencies between empirical fit and true dynamics in the eMLR calculation as the changes operate on scales that are too large to be filtered out by regression analysis.

7.2 Temporal sampling inconsistencies

The degree to which the use of a nominal time interval between sampling campaigns, an assumption we have made here, biases the estimated uptake rate is not clear. This depends on the spatial distribution of the data and on how the time interval is distributed spatially, i.e. how much each station influences the regressions.

The target North Atlantic average uptake rate in the simulation is $0.443\,\mathrm{Pg\,C\,yr^{-1}}$. This number is of course obtained from the knowledge that exactly 10 yr separate the measurements. Allowing for uncertainty in the timing of ± 2 yr (i.e. 8 or 12 yr), the uptake rate would vary from 0.52 to $0.34\,\mathrm{Pg\,C\,yr^{-1}}$. These values are close to the accuracy limits ($\pm 0.1\,\mathrm{Pg\,C\,yr^{-1}}$) on the uptake rate implied by the LSCOP criterion ($0.343\text{-}0.543\,\mathrm{Pg\,C\,yr^{-1}}$) indicating that a $0.1\,\mathrm{Pg\,C\,yr^{-1}}$ is about the same as a 2 yr error in timing in the North Atlantic.

Since most of the model formulae tend to produce North Atlantic uptake estimates that underestimate the true value (by 2–7.5 % on average, Fig. 7), if the characteristic time intervals of the dataset was smaller by 1 or 2 yr, a compensation in the rate calculation would occur. In contrast, the problem would become worse if the characteristic time interval were to be larger than 10 yr. Based on the noise-free calculations performed here, these considerations suggest that if a true inventory change can be approximated precisely, basin-scale eMLR-estimated uptake rates will remain within the desired accuracy of the true value if the bias in the characteristic time interval is smaller than about ± 2 yr.

The issue of temporally staggered samples is an important shortcoming of basin-scale eMLR that has yet to be addressed. Although interior DIC values can perhaps be adjusted to a nominal year (e.g. using the transient steady state concept, as used by Tanhua et al., 2007), it is not possible to do the same with all the tracers. This may result in possible inconsistencies as samples in different regions can be influenced by different modes or phases of natural variability patterns at different times, even within a sampling campaign.

7.3 Spatial sampling density

One important aspect affecting the quality of eMLR results is the spatial representativeness of the GLODAP and CLIVAR datasets. Ideally, datasets should measure approximately the same hydrographic regions with comparable relative sampling density, otherwise optimized empirical formulae may contain different explanatory variables. Emphasis on particular water masses or gradients may result not only from the presence or absence of data in the region, but also from the station density along hydrographic cruises.

As a consequence 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 type of overfitting, 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), structures due to regression misfit are more likely to cancel (Goodkin et al., 2011). A quantitative assessment of this effect is difficult with real data, however. The difficulty comes of course from the fact that the sampling grid varies, making point-by-point comparison of the regression misfit difficult without a form of interpolation. A detailed analysis at crossover stations may prove to be informative in that case.

For real data, the problem is also that the relevant systematic regression misfit should in principle be between the true anthropogenic signal and the empirical representation of it, not between the observed values, which are contaminated by natural variability, and the regression predictions. Nonetheless, a visual analysis of the geographical distribution of the residuals (calculated between the observed values and the regression predictions) associated with each station in our synthetic dataset indicates that residuals are not randomly distributed in space, nor are they totally uncorrelated between GLODAP and CLIVAR. Spatially coherent regions with residuals of the same sign (sometimes of similar magnitude) are generated (Plancherel, 2012).

Although model selection does not influence basin-scale estimates of the inventory changes very much, model selection is very important locally, affecting assessments of both the column and the layer inventories. The concepts of a balanced station coverage and of vertical continuity were used, in addition to statistical measures of fit, as guides for model selection in this study. Formally quantitative methodologies that account for these additional aspect as part of the eMLR calculation are desirable but are still lacking. It seems also conceivable to develop some criterion based on the pattern similarity of the residual field to help select appropriate regression models, to quantitatively exploit the systematic misfit cancelation effect. Incorporation of prior information in the derivation of the regression model could be used to limit the scope and variability of the regression structure in time and constrain the geographical coherence of the misfit.

In a few regions, multiple repeated cruises are available (e.g. OVIDE section in the Northeast Atlantic; Lherminier et al., 2007). Using all of these sections in the analysis will bias the dataset towards these particular regions. In such cases, it is of course best to use the one cruise track that is most representative of the nominal year used in the analysis (e.g. 1995, 2005). Due to the temporal data distribution, the basin-scale eMLR estimate of the carbon uptake represents a weighted average over a few years. High-frequency repeat sections provide a rare and valuable opportunity to evaluate the sensitivity of the final eMLR estimates to temporal data inconsistencies by replacing the temporally most representative section with the others. These repeat cruises can also be used to estimate the detection limit of eMLR directly from data.

7.4 Additional recommendations

As global eMLR implementations are currently being developed by the oceanographic community, a few additional points not addressed in this paper but relevant to the application of eMLR to real data are worthy of mention here. First, this study focused on the accuracy of eMLR. The related question of precision was only treated briefly in the theory section, and the influence of measurement errors and possible biases between data from different cruises remains to be addressed. Secondly, while working in smaller geographical regions will improve the regression fits, the size of regions should not be so small as to be prone to strong aliasing by variability of characteristic time scales shorter than the time scale inherent to the Repeat Hydrography program (about 10 yr). This issue is related to the systematic misfit cancelation effect as the risk of model overfitting (in the pragmatic sense of eMLR) is greater in highly dynamic regions. The risk of overfitting for analyses performed on sections, particularly if these are cut up in small pieces, is high.

Finally, this analysis was performed on depth layers, to mimic previous model-based assessments and for convenience (the model output is gridded to depth levels). An analysis performed on isoneutral surfaces instead of horizontal surfaces would likely perform better as property gradients on isoneutrals are smaller given that water masses mostly mix along these surfaces and because isoneutral outcrops naturally follow dynamical features. Isoneutral surfaces slope and cross the nutricline, however. Since the dominant mode of spatial variability of nutrient-type tracers in the ocean is due to their vertical distribution (Fukumori and Wunsch, 1991), there is a trade-off between the type of variability eMLR has to cope with between a mixing-dominated horizontal analysis and a biology-dominated isoneutral analysis. Both approaches are subject to the problem of variable end-member properties (i.e. secular trends), though. The influence of the nutricline on the analysis will depend on the location and density of the stations relative to the topography of the isoneutral surfaces. Solution of inverse problems, such as eMLR, are best when variables contained in the design matrix $\mathbf{Z}$ are independent. Unfortunately, oceanographic tracers tend to be highly correlated. If the vertical gradients projected on the isoneutrals becomes an important component of the isoneutral variance, it is possible that the problem of tracer colinearity may also be more important

along isoneutral than along depth horizons since roughly 30–50 % of the interior nutrient concentrations represent remineralized nutrients and the remineralization signal is highly correlated across nutrients (Anderson and Sarmiento, 1994). Nonetheless, optimizing tracer orthogonality, perhaps by using quasi-conservative tracers ($Si^\star$, $N^\star$, $C^\star$, PO, NO) or by adding dynamic tracers (potential vorticity, sea-surface height) should improve the conditioning of the problem, resulting in more appropriate regression fits.

8 Conclusions

Recasting the eMLR equations in the formalism of inverse problems allows for different application strategies for eMLR, including regression models that can change in time. This opens the conceptual possibility of systematically using empirical models that represent best-fit regressions that reflect the changing structure of the observational networks available. This perspective contrasts with the traditional approach that relies on model formulae that are fixed in time. The performance of these two approaches 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 datasets that mimic the spatial structure of the observed historical datasets.

Comparing eMLR results obtained by holding regression formulae fixed in time with results obtained by regression formulae that are allowed to change to reflect differences in dataset variance imposed by a redistribution of the oceanographic stations shows that more accurate results are possible when the structure of the empirical model fits is held constant in time. Given the working definitions of GLODAP and CLIVAR used here, this statement holds for basin-integrated estimates, layer-specific inventory change profiles and for maps of column inventory changes.

Comparison of idealized experiments in which the observational network is held fixed with realistic cases that incorporate the GLODAP to CLIVAR change in coverage indicates that best results are achieved when the GLODAP stations are used at all times. GLODAP results are superior to the CLIVAR results because GLODAP samples the whole North Atlantic more evenly than CLIVAR does, even if GLODAP has only half as many stations as CLIVAR. This results in empirical regression models for GLODAP that are more representative of the North Atlantic as a whole. In contrast, CLIVAR models tend to be more influenced by the subtropics because of the heavier station density there.

Holding the observational network fixed in time reduces the sensitivity of eMLR results to regression model selection relative to the case when the network changes in time. This is because the systematic misfit cancelation effect is less likely to operate when the station coverage changes in time, making regression model selection an even more important step when station coverage changes. Changing the model structure in time to better fit the observations induces changes in the signal-to-noise partitioning, de facto altering the working definitions of anthropogenic carbon and natural carbon imposed by the choice of regression model. Keeping the formula structure fixed increases the likelihood that the systematic biases inherent to using empirical representations of the true processes governing the distribution of anthropogenic carbon form in the same geographical regions and thus cancel during subtraction in the eMLR calculation.

As hydrographic station coverage is inherently sparse and changes between observational networks are significant relative to interannual variability and relative to the large-scale spatial variance patterns on horizontal layers, best-fit regression models can behave as if they were in fact overfitted; that is regionally over-specialized, yielding temporally inconsistent empirical definitions of the processes controlling tracer distribution on larger or longer time-scales. For this reason, simpler regression models, which may produce higher residual errors, may also yield better eMLR solutions as they make inherently fewer assumptions about the structure of the signal than more complicated models. More complex models may be driven towards regional fits at the expense of the broader picture.

Statistical fitness of the regressions, although helpful and necessary to some degree, is not a sufficient criterion for regression model selection in eMLR. Consideration of the spatial representativeness of the sampling network, vertical continuity of the selected regression formulae as justified by oceanographic knowledge and resiliency of the spatial structure of the residual patterns to temporal variability and changes in observational networks should be used as additional criteria to aid the model selection process and reduce systematic biases of eMLR results.

Most eMLR cases considered (most regression models and assumptions regarding the observational networks) can reproduce the simulated basin-integrated ocean carbon decadal inventory change within the threshold of acceptable uncertainty (10 %), as proposed in the LSCOP report (Bender et al., 2002). Analysis of layer-specific and column inventory changes indicate, however, that both the station distribution and the selection of regression models exert strong influences on eMLR's ability to recover the true signal locally. Both layer-specific and column 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 relatively good agreement with the true value. The depth range between 500 and 2700 m is particularly sensitive to model selection. In general, the subpolar convective region is the source of most of the difference between the true signal and the eMLR-inferred signals in the North Atlantic. This is a consequence of insufficient sampling in this region during

CLIVAR and also because of the strong variability in convective activity in the region, a type of variability that is not represented statistically in the data.

Implicit dynamical relationships between the size of the domain analyzed (or the density of samples in particular regions) and the time scales characteristic of that domain place limits on the ability of eMLR to account for natural variability. As eMLR relies on a statistical approach to filter out noise, for the residual field to be representative of natural variability, the spatial scales of the dominant modes of natural variability in the domain should be smaller than the size of the domain. Modes of natural variability similar in scale to the size of the domain cannot be discounted as noise by the spatial regressions used in eMLR. Large-scale natural variability patterns are treated instead as secular trends by the fitting process and ultimately contaminate the anthropogenic signal. Inadequate station density and inhomogenous sampling exacerbate these limitations.

Even if the best eMLR results obtained here are unable to fully account for the large-scale natural variability pattern, eMLR is able to remove a large fraction of it, despite our direct and somewhat naive horizontal (instead of isoneutral) analysis. The fact that eMLR produces relatively good results in the hydrographically complex and dynamic North Atlantic suggests that it is likely to perform well in other hydrographically simpler and less variable basins. Although further development and assessment of the method is necessary, particularly to address the issues of temporally variable covariances, full propagation of the errors, problem conditioning and temporal staggering of the samples, and even if inherent limitations exist imposed by the scale of the system in relation to the relevant modes of natural variability, the eMLR approach remains a viable candidate that can be used to exploit the many interior DIC measurements and evaluate the large-scale evolution of the ocean carbon sink and its rate of change independently from other techniques. Regional results should be interpreted with caution, however.

Appendix A

Spatio-temporal variance patterns in the synthetic data set

Given that regression analysis aims to explain the dominant variance patterns in a dataset, changes in the spatial and temporal patterns of variance can affect eMLR results by influencing regression model selection. Variance variability can arise either from temporal variability or by altering the sampling grid, which acts by weighting certain regions differently in the dataset. The structure and quality of linear regressions vary depending on whether the analyses are performed on sections, on regions, or on isopycnals such that the regression models used in the eMLR context are ad hoc. This section contrasts the spatial variance patterns captured by the GLODAP and CLIVAR sampling networks and discusses the seasonal to interannual changes of these spatial patterns in the synthetic dataset used in this study.

Figure A1 shows the seasonal evolution of vertical profiles of the standard deviation in the synthetic North Atlantic GLODAP dataset for year 1995 for 8 variables. The standard deviation is calculated horizontally and independently for each month and each model layer. A parallel analysis using the CLIVAR sampling grid shows similar broad-scale patterns, although with slightly different magnitudes owing to the different emphasis put on the Labrador Sea and the Eastern Tropical Atlantic between the two sampling networks. The variables exhibit different zones of low or high variance (Fig. A1), indicating a priori the role each tracer will take in the regression models as a function of depth and highlighting the value of each variable as a tracer for each water masses.

The seasonal evolution of variance profiles reflects the mechanisms of water mass formation, gas exchange and ecological succession in the basin. The magnitude of the seasonal cycle of the standard deviation is typically 10 to 15 % in the upper 200 m for the nutrients (O_2, AOU, NO_3, PO_4, Si), and 5 % for θ, S, Alk and DIC. Seasonality is small below 200 m (< 1–2 %). Nutrients show large variances in late summer and fall in the top 150 m and relatively smaller standard deviations in winter and spring (Fig. A1), consistent with the development of the North Atlantic Bloom (Henson et al., 2009). Temperature shows a maximum variance in spring and summer when the subtropical-subpolar gradients are strongest. The variance of salinity is small in summer and is large in winter, reflecting sea-ice dynamics in the northern subpolar region. Seasonality of the variance is associated with a seasonal cycle in the misfit error of linear regression models and in the eMLR results.

Relative to the basin-scale horizontal variance in the dataset, 1995 to 2005 variance changes in the vertical profiles are small. These changes are typically less than 3 % above 500 m and less than 1 % below that depth. These changes reflect processes such as water mass reorganization, gyre wobble, thermocline oscillation, frontal shifts, etc. Although the level of variance on horizontal slices in the data are relatively constant, this is not to say that point-by-point differences in tracer values or concentrations do not routinely exceed the standard deviation calculated over the whole layer. In fact, point-by-point differences between July 1995 and 2005 for the North Atlantic can be as high as 50–100 % in specific regions (East Greenland Current, Labrador Sea, across the North Atlantic Current, near the equatorial boundary of the subtropical gyre). The relative constancy of the dataset variance in time sampled from a constant observational network suggests that the point-by-point changes are not a priori systematic enough as to greatly bias the large-scale representativeness of a given sampling grid: the GLODAP or CLIVAR sets of stations would measure features in the same approximate proportions in 1995 and in 2005.

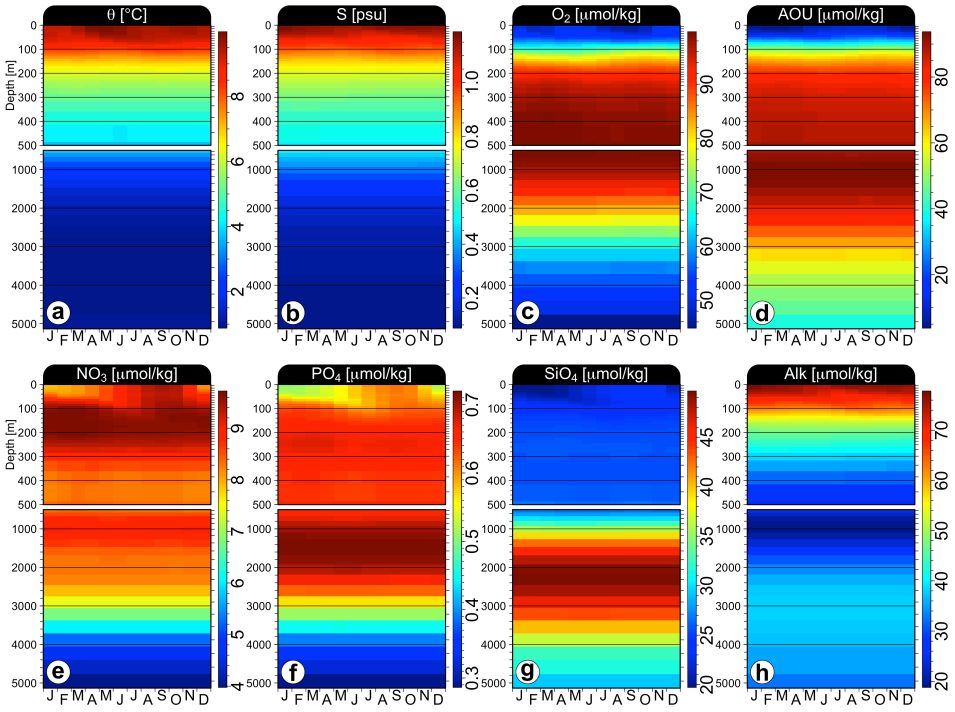

Fig. A1. Monthly vertical profiles of the horizontal spatial standard deviation, from January to December, expressed in **(a)** °C, **(b)** psu or **(c–h)** µmol kg^{-1}, for the hydrographic variables used in this study for the year 1995 as sampled on the GLODAP grid. Tick marks to the right of the main panels show the vertical position of the vertical layers in the circulation model.

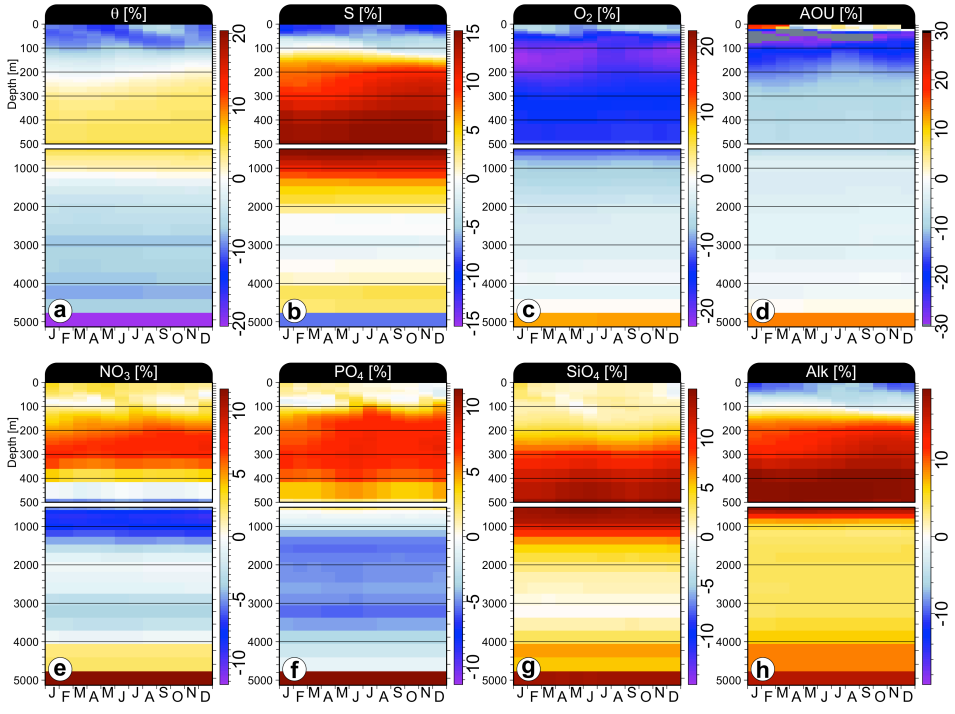

Fig. A2. Monthly vertical profile of the month-by-month relative changes of the horizontal spatial standard deviation, from January to December, expressed in percent relative to the 1995 values, between the synthetic 2005 CLIVAR dataset and the 1995 GLODAP dataset for the variables used in this study. Tick marks to the right of the main panels show the vertical position of the layers in the circulation model.

The GLODAP and CLIVAR sampling grids emphasize hydrographic features differently because of their variable spatial sampling densities (Fig. 1). Calculated differences between the basin-scale variances of the GLODAP and CLIVAR datasets show typical standard deviation differences of order 10 % between the two observational networks. These differences also exhibit vertical patterns clearly different from changes induced by natural variability or either seasonal or interannual time-scales (Fig. A2). Interannual variability and variations in the sampling grid both alter the dataset variance patterns and affect misfit error but interannual variability is secondary to the variance changes imposed by changing sampling network between GLODAP 1995 and CLIVAR 2005.

Supplementary material related to this article is available online at: http://www.biogeosciences.net/10/4801/2013/bg-10-4801-2013-supplement.pdf.

Acknowledgements. This study was supported by the National Science Foundation under Grant No. OCE-0727170 and OCE-0327189 and by BP and Ford Motor Company through the Carbon Mitigation Initiative at Princeton University. We thank J. Dunne for providing model output. Y. Plancherel thanks the Oxford Martin School for support during completion of this research. The contributions of K. B. Rodgers, A. R. Jacobson, and R. M. Key were supported through NASA award NNX09AI13G. Additionally, KBR was supported through NOAA awards NA17RJ2612 and NA08OAR4320752, which includes support through the NOAA Office of Climate Observations (OCO). R. M. Key acknowledges additional support from NOAA awards NA08OAR4310820 and NA08OAR4320752. Lastly, we would like to thank Are Olsen and three other reviewers for their helpful and constructive comments.

Edited by: C. Heinze

References

Anderson, L. A. and Sarmiento, J. L. S.: Redfield ratios of remineralization determined by nutrient data analysis, Global Biogeochem. Cy., 8, 65–88, 1994.

Bates, N. R.: Interannual variability of oceanic CO_2 and biogeochemical properties in the Western North Atlantic subtropical gyre, Deep-Sea Res. Pt. II, 48, 1507–1528, 2001.

Bates, N., Michaels, A., and Knap, A.: Seasonal and interannual variability of oceanic carbon dioxide species at the U.S. JGOFS Bermuda Atlantic Time-series Study (BATS) site, Deep-Sea Res. Pt. II, 43, 347–383, 1996.

Bender, M., Doney, S., R. A., Feely, Fung, I., Gruber, N., Harrison, D. E., Keeling, R., Moore, J. K., Sarmiento, J., Sarachik, E., Stephens, B. Takahashi, T., Tans, P., and Wanninkhof, R.: A large-scale CO_2 observing plan: In situ oceans and atmosphere (LSCOP), OAR Special Report NTIS: PB2003-00377, NOAA/OAR/PMEL, Seattle, WA, 2002.

Brewer, P. G., Goyet, C., and Friederich, G.: Direct observation of the oceanic CO_2 increase revisited, P. Natl. Acad. Sci. USA, 94, 8308–8313, 1997.

Broecker, W. S.: "NO", a conservative water-mass tracer, Earth Planet. Sc. Let., 23, 100–107, 1974.

Broecker, W. S., Takahashi, T., and Takahashi, T.: Sources And Flow Patterns Of Deep-Ocean Waters As Deduced From Potential Temperature, Salinity, And Initial Phosphate Concentration, J. Geophys. Res.-Oceans, 90, 6925–6939, 1985.

Burnham, K. and Anderson, D.: Model selection and multimodel inference: a practical information-theoretic approach, Springer-Verlag, New York, 2nd Edn., 1998.

Caldeira, K. and Duffy, P. B.: The role of the Southern Ocean in uptake and storage of anthropogenic carbon dioxide, Science, 287, 620–622, 2000.

Corbiere, A., Metzl, N., Reverdin, G., Brunet, C., and Takahash.T.: Interannual and decadal variability of the oceanic carbon sink in the North Atlantic subtropical gyre, Tellus B, 59, 168–178, 2007

Cromwell, D.: Temporal and spatial characteristics of sea surface height variability in the North Atlantic Ocean, Ocean Sci., 2, 147–159, doi:10.5194/os-2-147-2006, 2006.

Curry, R., McCartney, M. S., and Joyce, T. M.: Oceanic transport of subpolar climate signals to mid-depth subtropical waters, Nature, 391, 575–577, 1998.

Dunne, J. P., Armstrong, R. A., Gnanadesikan, A., and Sarmiento, J. L.: Empirical and mechanistic models for the particle export ratio, Global Biogeochem. Cy., 19, GB4026, doi:10.1029/2004GB002390, 2005.

Dunne, J., Sarmiento, J., and Gnanadesikan, A.: A synthesis of global particle export fromthe surface ocean and cycling through the ocean interior and on the seafloor, Global Biogeochem. Cy., 21, GB4006, doi:10.1029/2006GB002907, 2007.

Dunne, J., Gnanadesikan, A., and Sarmiento, J. L.: Coupling between the C,N, P, Fe, Si, and Ca, and lithogenic cycles in a global biogeochemical and ecological model., Abstract ID:1166, Ocean Sciences Meeting, Orlando, 2–7 March 2008.

Dunne, J., Gnanadesikan, A., Sarmiento, J. L., and Slater, R.: Technical description of the prototype version (v0) of Tracers of Phytoplankton with Allomeric Zooplankton (TOPAZ) ocean biogeochemical model as used in the Princeton IFMIP model, Supplement to Sarmiento et al. 2010, Biogeosciences, 7, 3593–3624, 2010,
http://www.biogeosciences.net/7/3593/2010/.

Friis, K., Kortzinger, A., J., P., and Wallace, D. W. R.: On the temporal increase of anthropogenic CO_2 in the subpolar North Atlantic, Deep-Sea Res. Pt. I, 52, 681–698, 2005.

Fukumori, I. and Wunsch, C.: Efficient representation of the North Altantic hydrographic and chemical distributions, Prog. Oceanogr., 27, 111–195, 1991.

Gnanadesikan, A., Dixon, K. W., Griffies, S. M., Balaji, V., Barreiro, M., Beesley, J. A., Cooke, W., Delworth, T. L., Gerdes, R., Harrison, M. J., Held, I. M., Hurlin, W. J., Lee, H.-C., Liang, Z., Nong, G., Pacanowski, R. C., Rosati, A., Russell, J., Samuels, B. L., Song, Q., Spelman, M. J., Stouffer, R. J., Sweeney, C. O., Vecchi, G., Winton, M., Wittenberg, A. T., Zeng, F., Zhang, R., and Dunne, J. P.: GFDL's CM2 global coupled climate models. Part II: The baseline ocean simulation, J. Climate, 19, 675–697, 2006.

Goodkin, N., Levine, N., Doney, S., and Wanninkhof, R.: Impacts of temporal CO_2 and climate trends on the detection of ocean anthropogenic CO_2 accumulation, Global Biogeochem. Cy., 25, GB3023, doi:10.1029/2010GB004009, 2011.

Griffies, S. M., Harrison, M., Pacanowski, R. C., and Rosati, A.: A technical guide to MOM4, Technical report, GFDL Ocean Group, Princeton, 2004.

Griffies, S. M., Gnanadesikan, A., Dixon, K. W., Dunne, J. P., Gerdes, R., Harrison, M. J., Rosati, A., Russell, J. L., Samuels, B. L., Spelman, M. J., Winton, M., and Zhang, R.: Formulation of an ocean model for global climate simulations, Ocean Sci., 1, 45–79, doi:10.5194/os-1-45-2005, 2005.

Griffies, S., Biastoch, A., Boening, C., Bryan, F., Danabasoglu, G., Chassignet, E. P., England, M. H., Gerdes, R., Haak, H., Hallberg, R. W., Hazeleger, W., Jungclaus, J., Large, W. G., Madec, G., Pirani, A., Samuels, B. L., Scheinert, M., Sen Gupta, A., Severijns, C. A., Simmons, H. L., Treguier, A. M., Winton, M., Yeager, S., and Yin, J.: Coordinated ocean-ice reference experiments (COREs), Ocean Model., 26, 1–46, 2009.

Gruber, N., Gloor, M, Mikaloff Fletcher, S. E., Doney, S. C., Dutkiewicz, S., Follows, M. J., Germer, M. Jacobson, A. R., Joos, F., Lindsay, K., Menemenlis, D., Mouchet, A., Mueller, S. A., Sarmiento, J. L., and Takahashi, T.: Oceanic sources, sinks, and transport of atmospheric CO_2, Global Biogeochem. Cy., 23, GB1005, doi:10.1029/2008GB003349, 2009.

Hartigan, J. and Wong, M.: A K-means clustering algorithm, Appl. Statist., 28, 100–108, 1979.

Henson, S., Dunne, J., and Sarmiento, J.: Decadal variability in North Atlantic phytoplankton blooms, J. Geophys. Res., 114, C04013, doi:10.1029/2008JC005139, 2009.

Henson, S. A., Sarmiento, J. L., Dunne, J. P., Bopp, L., Lima, I., Doney, S. C., John, J., and Beaulieu, C.: Detection of anthropogenic climate change in satellite records of ocean chlorophyll and productivity, Biogeosciences, 7, 621–640, doi:10.5194/bg-7-621-2010, 2010.

Keeling, R. F.: Comment on "The ocean sink for anthropogenic CO_2", Science, 308, 1743c, 2005.

Key, R. M., Kozyr, A., Sabine, C. L., Lee, K., Wanninkhof, R., Bullister, J. L., Feely, R. A., Millero, F. J., Mordy, C., and Peng, T.-H.: A global ocean carbon climatology: Results from Global Data Analysis Project (GLODAP), Global Biogeochem. Cy., 18, GB4031, doi:10.1029/2004GB002247, 2004.

Khatiwala, S., Primeau, F., and Hall, T.: Reconstruction of the history of anthropogenic CO_2 concentrations in the ocean, Nature, 462, 346–350, 2009.

Large, W. G. and Yeager, S. G.: Diurnal to decadal global forcing for ocean and sea-ice models: the datasets and flux climatologies, Tech. Rep. TN-460+STR, National Center for Atmospheric Research, Boulder, Colorado, 2004.

Large, W. G. and Yeager, S. G.: The global climatology of an internanually varying air-sea flux dataset, Clim. Dynam., 33, 341–364, 2009.

Lee, K., Choi, S. D., Park, G. H., Wanninkhof R., Peng, T.-H., Key, R. M., Sabine, C. L., Feely, R. A., Bullister, J. L., Millero, F. J., and Kozyr, A.: An updated anthropogenic CO_2 inventory in the Atlantic ocean, Global Biogeochem. Cy., 17, 1116, doi:10.1029/2003GB002067, 2003.

Le Quéré, C., Rodenbeck, C, Buitenhuis, E. T., Conway, T. J., Langenfelds, R., Gomez, A., Labuschagne, C., Ramoney, M., Nakazawa, T., Metzl, N., Gillett, N., and Heinmann, M.: Saturation of the Southern Ocean CO_2 Sink Due to Recent Climate Change, Science, 316, 1735–1738, 2007.

Levine, N., Doney, S. C., Wanninkhof, R., Lindsay, K., and Fung, I.: Impact of ocean carbon system variability on the detection of temporal increases in anthropogenic CO_2, J. Geophys. Res., 113, C03019, doi:10.1029/2007JC004153, 2008.

Lherminier, P., Mercier, H., Gourcuff, C., Alvarez, M., Bacon, S., and Kermabon, C.: Transports across the 2002 Greeland-Portugal Ovide section and comparsion with 1997, J. Geophys. Res., 112, C07003, doi:10.1029/2006JC003716, 2007.

Lo Monaco, C., Metzl, N., Poisson, A., Brunet, C., and Shauer, B.: Anthropogenic CO_2 in the Southern Ocean: Distribution and inventory at the Indian-Atlantic boundary (World Ocean Circulation Experiment line I6). J. Geophys. Res., 110, C06010, doi:10.1029/2004JC002643, 2005a.

Lo Monaco, C., Goyet, C., Metzl, N., Poisson, A., and Touratier, F.: Distribution and inventory of anthropogenic CO_2 in the Southern Ocean: comparson of three data-based methods, J. Geophys. Res., 110, C09S02, doi:10.1029/2004JC002571, 2005b.

Mackas, D. L., Denman, K. L., and Bennett, A. F.: Least-square multiple tracer analysis of water mass composition, J. Geophys. Res.-Oceans, 92, 2907–2918, 1987.

Matear, R. and McNeil, B.: Decadal accumulation of anthropogenic CO_2 in the Southern Ocean: A comparison of CFC-age derived estimates to multiple-linear regression estimates, Global Biogeochem. Cy., 17, 1113, doi:10.1029/2003GB002089, 2003.

McKinley, G. and Fay, A. R., Takahashi, T., and Metzl, N.: Convergence of atmostpheric and North Atlantic carbon dioxide trends on multidecadal timescales, Nat. Geosci., 4, 606–610, 2011.

Mikaloff-Fletcher, S. E. and Gruber, N. and Jacobson, A. R., Doney, S. C., Dutkiewicz, S., Gerber, M., Follows, M., Joos, F., Lindsay, K., Menemenlis, D., Mouchet, A., Mueller, S. A., and Sarmiento J. L.: Inverse estimates of anthropogenic CO_2 uptake, transport, and storage by the ocean, Global Biogeochem. Cy., 20, GB2002, doi:10.1029/2005GB002530, 2006.

Plancherel, Y.: A study of the ocean's water masses using data and models. Ph.D. Thesis, Princeton University, Princeton, USA, 2012.

Rio, M.-H., Guinehut, S., and Larnicol, G.: New CNES-CLS09 global mean dynamic topography computed from the combination of GRACE data, altimetry and in situ measurements, J. Geophys. Res., 116, C07018, doi:10.1029/2010JC006505, 2011.

Rodgers, K. B., Key, R., Gnanadesikan, A., Sarmiento, J. L., Aumont, O., Bopp, L., Doney, S. C., Dunne, J. P., Glover, D. M., Ishida, A., Ishii, M., Jacobson, A. R., Lo Monaco, C., Maier-Reimer, E., Mercier, H., Metzl, N., Perez, F. F., Rios A. F., Wawnninkhof, R., Wetzel, P., Winn, C., and Yamanaka, Y.: Using altimetry to help explain patchy changes in hydrographic carbon measurments, J. Geophys. Res., 114, C09013, doi:10.1029/2008JC005183, 2009.

Sabine, C. L. and Feely, R. A. and Gruber, N., Key, R. M., Lee, K., Bullister, J. L., Wanninkhof, R., Wong, C. S., Wallace, D. W. R., Tilbrook, B., Millero, F. J., Peng, T.-H., Kozyr, A., Ono, T., and Rios, A. F.: The oceanic sink for anthropogenic CO_2, Science, 305 , 367–371, 2004.

Sabine, C., Feely, R., Millero, F., Dickson, A. G., Langdon, C., Mecking, S. and Greeley, D.: Decadal changes in Pacific carbon, J. Geophys. Res., 113, C07021, doi:10.1029/2007JC004577,

2008.

Sarmiento, J. L., Slater, R. D., Dunne, J., Gnanadesikan, A., and Hiscock, M. R.: Efficiency of small scale carbon mitigation by patch iron fertilization, Biogeosciences, 7, 3593–3624, doi:10.5194/bg-7-3593-2010, 2010.

Schuster, U. and Watson, A. J.: A variable and decreasing sink for atmospheric CO_2 in the North Atlantic, J. Geophys. Res., 112, C11006, doi:10.1029/2006JC003941, 2007.

Sonnerup, R. E., Quay, P. D., and McNichol, A. P.: The Indian Ocean ^{13}C Suess effect, Global Biogeochem. Cy., 14, 903–916, 2000.

Steinfeldt, R., Rhein, M., Bullister, J., and Tanhua, T.: Inventory changes in anthropogenic carbon from 1997-2003 in the Atlantic Ocean between 20° S and 65° N, Global Biogeochem. Cy., 23, GB3010, doi:10.1029/2008GB003311, 2009.

Takahashi, T, Sutherland, S. C., Sweeney, C., Poisson, A., Metzl, N., Tilbrook, B., Bates, N., Wanninkhof, R., Reely, R. A., Sabine, C., Olafsson, J., and Nojiri, Y.: Global sea-air CO_2 flux based on climatological surface ocean pCO_2, and seasonal biological and temperature effects, Deep-Sea Res. Pt. II, 49, 1601–1622, 2002.

Takahashi, T, Sutherland, S. C., Wanninkhof, R., Sweeney, C., Feely, R. A., Chipman, D. W., Hales, B., Friederich, G., Chavez, F., Sabine, C., Watson A., Bakker, D. C. E., Schuster, U., Metzl, N., Yoshikawa-Inoue, H., Ishii, M., Midorikawa, T., Nojiri, Y., Koertzinger, A., Steinhoff, T., Hoppema, M., Olafsson, J., Arnarson, T. S., Tilbrook, B., Johannessen, T., Olsen, A., Bellerby, R., Wong, C. S., Delille B., Bates, N. R., and de Baar, H. J. W.: Climatological mean and decadal change in surface ocean pCO_2, and net sea-air CO_2 flux over the global oceans, Deep-Sea Res. Pt. II, 56, 554–577, 2009.

Tanhua, T., Koertzinger, A., Friis, K., Waugh, D. W., and Wallace D. W. R.: An estimate of anthropogenic CO_2 inventory from decadal changes in oceanic carbon content, P. Natl. Acad. Sci. USA, 104, 3037–3042, 2007.

Tarantola, A.: Inverse problem theory and methods for model parameter estimation, SIAM, Philadelphia, PA, 2005.

Wallace, D. W. R.: Monitoring Global Ocean Carbon Inventories, Tech. rep., Ocean Observing System Development Panel, Texas A&M University, College Station, TX, 1995.

Wanninkhof, R., Doney, S. C., Bullister, J., Levine, N. M., Warner, M., and Gruber, N.: Detecting anthropogenic CO_2 changes in the interior Atlantic ocean between 1989–2005, J. Geophys. Res., 115, C11028, doi:10.1029/2010JC006251, 2010.

Waugh, D. W. and Hall, T. M., McNeil, B. I., Key, R. M., and Matear, R. J.: Anthropogenic CO_2 in the oceans estimated using transit time distributions, Tellus, 58B, 376–389, 2006.

Wetzel, P., Winguth, A., and Maier-Reimer, E.: Sea-to-air CO_2 flux from 1948 to 2003: A model study, Global Biogeochem. Cy.,19, GB2005, doi:10.1029/2004GB002339, 2005.

Winn, C. D., Li, Y. H., Mackenzie, F. T., and Karl, D. M.: Rising surface ocean dissolved inorganic carbon at the Hawaii Ocean Time-series site, Mar. Chem., 60, 33–47, 1998.

Winton, M.: A reformulated three-layer sea ice model, Journal of atmospheric and oceanic technology, 17, 525–531, 2000.