# High-resolution climatology of the northeast Atlantic using Data-Interpolating Variational Analysis (Diva)

C. Troupin,[1] F. Machín,[2] M. Ouberdous,[1] D. Sirjacobs,[1] A. Barth,[1] and J.-M. Beckers[1]

[1]   Numerous climatologies are available at different resolutions and cover various parts of the global ocean. Most of them have a resolution too low to represent suitably regional processes and the methods for their construction are not able to take into account the influence of physical effects (topographic constraints, boundary conditions, advection, etc.). A high-resolution atlas for temperature and salinity is developed for the northeast Atlantic Ocean on 33 depth levels. The originality of this climatology is twofold: (1) For the data set, data are collected on all major databases and aggregated to lead to an original data collection without duplicates, richer than the World Ocean Database 2005, for the same region of interest. (2) For the method, climatological fields are constructed using the variational method Data-Interpolating Variational Analysis. The formulation of the latter allows the consideration of coastlines and bottom topography, and has a numerical cost almost independent on the number of observations. Moreover, only a few parameters, determined in an objective way, are necessary to perform an analysis. The results show overall good agreement with the widely used World Ocean Atlas, but also reveal significant improvements in coastal areas. Error maps are generated according to different theories and emphasize the importance of data coverage for the creation of such climatological fields. Automatic outlier detection is performed, and its effects on the analysis are examined. The method presented here is very general and not dependent on the region, hence it may be applied for creating other regional atlas in different zones of the global ocean.

## 1.   Introduction

[2]  Gridded fields of hydrographic properties constitute a highly demanded product in a wide range of applications in oceanography, going from simple visualization purposes to initialization of numerical model and quality check of newly acquired data. A climatology is defined as a set of such gridded fields, computed at different levels, either in the global ocean or in one subregion.

[3]  Many techniques allow the creation of climatologies using in situ measurements. They differ in the way data are treated (individual or groups), the choice of vertical coordinate system (pressure or density), the a posteriori filters applied to the results and, above all, the mathematical background for solving the gridding problem.

[4]  Generally, two main approaches have to be distinguished for data gridding: (1) the interpolation: it implies a strict passage of the solution through the points of data, but it is obviously not adapted for the creation of climatologies, given the nature of the considered data and (2) the approximation (or analysis) provides a solution that does not necessarily contains all the data points: the reconstructed field is forced to be relatively close to the data points, or put in other words, the separation between data and field is adjusted. Approximation techniques are particularly well adapted to oceanographic measurements, since the latter are contaminated by several sources of errors or noise with respect to climatological values, making the use of strict interpolation obsolete.

[5]  The most popular global climatologies are the successive versions of the World Ocean Atlas (WOA) [*Levitus*, 1982; *Levitus and Boyer*, 1994; *Stephens et al.*, 2002; *Locarnini et al.*, 2006] which provide 1° by 1° fields on 33 isobaric levels. Although they are still frequently used for initializing numerical models [e.g., *Chassignet et al.*, 2003; *Marchesiello et al.*, 2003], they suffer from a lack of resolution, specially in the coastal zones.

[6]  A 1/4° WOA climatology was recently developed [*Boyer et al.*, 2005], yet coastal zones are still not adequately covered. This is a paradox characteristic of many

---

[1]GeoHydrodynamics and Environment Research, AGO, University of Liège, Liège, Belgium.
[2]Facultad de Ciencias del Mar, Departamento de Física, Universidad de Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain.
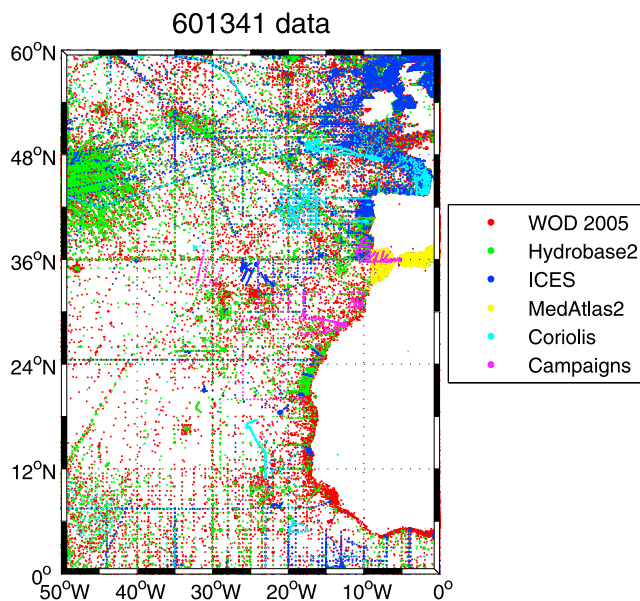
**Figure 1.** Localization of the profiles by data set. Note that WOD 2005 data are not always visible due to overlay with other databases.

climatologies, since coastal areas are generally more surveyed than the open ocean (see Figures 1 and 2). Another concern of this climatology is the application of smoothing filters on the resulting fields, in order to remove the noise coming from limited data coverage and small smoothing radius [*Chang and Chao*, 2000].

[7] *Gouretski and Koltermann* [2004] prepared the World Ocean Circulation Experiment (WOCE) Global Hydrographic Climatology (WGHC) on 45 standard levels with a 0.5 by 0.5° resolution. They worked with a combined data set including WOD98 and WOCE. Observations are separated into two groups: reference data, composed of high-quality observations (around 20000 profiles), and historical data (around 1,000,000 profiles). The first group serves as a reference for validating historical data and treating systematic biases. Potential density surface are used for averaging on deep layers, while isobaric surfaces are preferred for upper layers. The correlation function is a Gaussian form with a decorrelation length scale $R$. To account for smaller-scale processes in shallow region, $R$ is computed as a function of the distance between a given point and the coast. The problem of basin separation is solved by applying Optimal Interpolation (OI) individually on different sub-regions of the Ocean, the coherence between the solution being assured by zones of overlap. Their method provides tighter $T$-$S$ sequences compared with WOA01, but the OI method does not allow them to consider the effect of coastlines and to have sufficient resolution. Moreover, by the time their climatology was produced, they could not consider important data sources, such as Coriolis or MEDAR/MEDATLAS II.

[8] On a more regional scale, *Lozier et al.* [1995] proposed a North Atlantic climatology based on a database enriched with respect to WOA. As the data coverage was improved and the size of the domain reduced, they were able to use a smaller smoothing radius in the interpolation, thus

representing features absent from the WOA. Another enhancement is the use of isopycnic levels instead of isobaric ones: this allowed a better conservation of water mass structures. Despite these modifications, their results also suffer from the coastal issue, since observations located in regions shallower than 200 m were removed from their data set.

[9] A recurring drawback of many methods is the use of an isotropic correlation length in horizontal planes, while numerous physical processes often have different characteristics according to the direction (e.g., cross-shore versus alongshore variations). *Kim et al.* [2007] avoid this drawback by applying a two-step method to produce 5 km by 5 km gridded fields off the U.S. West Coast: (1) the data are averaged on a high-resolution sparse grid and (2) the averages are mapped on a regular grid using Objective Analysis (OA) [*Gandin*, 1965] with anisotropic correlation length. Nevertheless, the data binning prior to their analysis may constitute a loss of the information contained in the data.

[10] *Ridgway et al.* [2002] use locally weighted regression (also called loess mapping) [*Cleveland and Devlin*, 1988] on a combination of data sets for producing gridded fields around Australia. Their technique provides solutions to the referred problems, as it allows for anisotropy and topographical effect considerations. The resulting fields have a resolution that depends on the data coverage. The bias induced by the irregular temporal data coverage is removed by fitting the data to annual and semiannual harmonic components. The method requires a large number of adjustable parameters, hence making the analysis less objective and more difficult to adapt to a particular region. The authors also reported problems related to spatial sampling, such as instability in regions void of data, unrealistic temperature inversion, or effect of data clusters.

[11] Eventually, none of the climatologies that were described provides the user with complementary error fields,
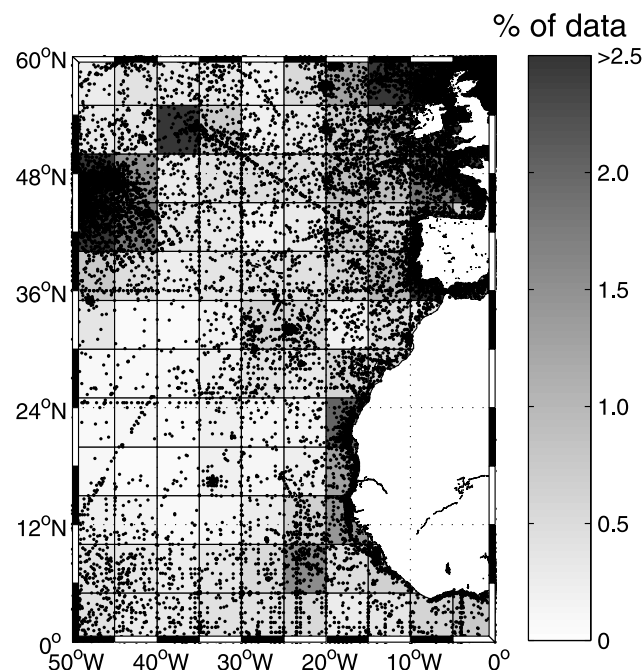


**Figure 2.** Relative data density on a 5° by 5° grid.

**Table 1.** Data Sources[a]

| Source | URL | Number of Profiles | |
|--------|-----|:---:|:---:|
| | | *T* | *S* |
| WOD05[b] | http://www.nodc.noaa.gov/OC5/WOD05/pr_wod05.html | 304,714 | 303,103 |
| HydroBase2 | http://www.whoi.edu/science/PO/hydrobase/ | 235,016 | 235,016 |
| ICES[c] | http://www.ices.dk/Ocean/data.asp | 45,552 | 45,002 |
| MEDATLAS2 | http://www.ifremer.fr/medar/cdrom_database.htm | 9413 | 9425 |
| Coriolis | http://www.coriolis.eu.org | 4367 | 4358 |
| Local Campaigns | – | 2330 | 2321 |
| Total | | 601,392 | 599,225 |

[a]*T* and *S* indicate the number of profiles before quality control and duplicate elimination for temperature and salinity, respectively.
[b]World Ocean Database 2005.
[c]International Council for the Exploration of the Sea.

except the WGHC [*Gouretski and Koltermann*, 2004]: they computed an absolute formal error as the product of the property variance and the relative error of the objective analysis.

[12] A different approach to the gridding problem, called process-convolution method (PCM) [e.g., *Higdon et al.*, 1998; *Swall*, 1999] has been recently developed and applied to hydrographic fields. The principle of the method is to represent a dependent process by the convolution of simple independent processes, using a kernel that varies with space and time, allowing for flexibility in modeling the spatial dependence [*Calder and Cressie*, 2007]. Therefore the PCM has the capacity of accounting for the temporal component of the data, which was systematically ignored in the previous approaches.

[13] *Higdon* [1998] applied this method to the temperature in the North Atlantic Ocean. As very few time series are available, the temporal dependencies of the kernel are not obvious to detect, hence they relied on standard variogram structures to fill in the gap. Because the numerical cost is important, the kernel was only determined at a limited number of locations. *Lemos and Sansó* [2006] worked on the temperature variability in the Portugal Current System using a similar approach. They employed a Gaussian kernel with horizontal and vertical standard deviations of 2° and 60 m, respectively, but the choice of these values for the parameters is not related to the data or to physical processes. Moreover the horizontal resolution was limited to 2° because of numerical cost.

[14] More recently, North Atlantic SST was analyzed by *Lemos and Sansó* [2009]. The kernel they applied allows the consideration of anisotropic and nonstationary fields. The resulting fields have the same large scale patterns than the WOA, but presents a better distinction of water masses. The estimation of the kernel parameters requires a substantial increase of the computational cost. This can constitute a temporary obstacle to the generation of a complete climatology (i.e. applying the same procedure to several variables at various depth levels).

[15] Having reviewed different methods, the objective is settled to the construction of a regional climatology with a gridding method that fulfills the following conditions: (1) deals with a great number of individual data, without needing to work with averaged values or data bins; (2) takes into account coastline and topography effects in a natural way, without additional parametrization; (3) generates

coherent error maps; and (4) relies on a limited number of parameters, estimated in an objective way. In addition, the technique implemented shall have reasonable computational time, both for analyzed and error fields, so that it can be run routinely on several depth levels and time periods without requiring huge computational resources.

[16] The selected approach is called Diva (Data-Interpolation Variational Analysis), an implementation of the Variational Inverse Method (VIM) [*Brasseur and Haus*, 1991; *Brasseur*, 1994], since this method satisfies all our requirements. It was successfully applied to the Mediterranean Sea for physical [*Brasseur et al.*, 1996] and biological variables [*Karafistan et al.*, 2002].

[17] The methodology presented in this paper is suitable to any region of the global ocean. For the sake of simplicity, we limit ourselves to a subregion of the Atlantic Ocean. This zone constitutes a pertinent choice for the application of Diva, since it gives the opportunity to deal with complex systems, such as the NW Africa upwelling system, the Strait of Gibraltar, the presence of Archipelagos (Canary, Madeira, Azores etc.) and several sea mounts or the plume of Amazon River.

[18] The paper is organized as follows: section 2 describes the preparation of the data and the preprocessing steps; section 3 gives an overview of Diva theory and of the selection of parameters for producing the climatology; section 4 is dedicated to the analyzed and error fields, while section 5 emphasizes the improvements brought by the method and the new data set; conclusions and future works are the subject of section 6.

## 2. Data

[19] As data coverage is an essential factor in the climatology quality, we strive to create the most complete database for our region of sake.

### 2.1. Sources

[20] Data were gathered from several databases (Table 1) including only CTD and bottle measurements. XBT data are known to exhibit systematic error due to incorrect fall rate [*Hanawa and Yasuda*, 1992]. Although correcting equations are available [*Hanawa et al.*, 1994], the decision to stick only to CTD and bottles was preferred, at least for the present version of the climatology. Moreover, it is necessary to know which type of XBT is considered (e.g. T4, T5, etc.) to apply the suitable correction. Unfortunately, the XBT

**Table 2.** Number of Profiles and Duplicates by Database

| Database | Duplicates | Percent | After Duplicate Removal |
|---|---|---|---|
| WOD05 | – | – | 304,714 |
| HydroBase2 | 225,341 | 95.9 | 9675 |
| ICES | 34,388 | 75.5 | 11164 |
| MEDATLAS2 | 7375 | 78.4 | 2038 |
| Coriolis | 3009 | 68.9 | 1358 |
| Local Campaigns | 836 | 35.9 | 1494 |
| Total | | | 330,443 |

type is rarely directly available. Nevertheless, conclusions drawn from the present work are still applicable to data sets made up of data types other than CTD and bottles.

[21] Within each individual database, when previously quality-controlled data were available, they were preferred to raw data. Figure 1 indicates the locations of the profiles for each database: coastal areas are particularly well covered, mostly near France, Spain and England. Mediterranean Sea and Gibraltar Strait benefit from a high data density, thanks to the MEDAR/MEDATLAS II project. The northwestern part of the domain (Newfoundland and Labrador shelves) is also well surveyed, as it is the zone where mixing between Gulf Stream and Labrador Current takes place. The least covered zone is by far the central part of the subtropical gyre, as only 7.9 percent of the data are located between 30–50°W and 10–30°N, with an average of about 600 profiles per 5°W box (Figure 2).

## 2.2. Preprocessing

### 2.2.1. Duplicate Detection

[22] The first step to obtain our database is the detection and removal of data duplicates. Otherwise, data analysis would be biased in places where two (or more) identical data are present, as the data weight at these locations would be artificially multiplied. The criterion to detect duplicates is the following: two profiles measured at the same time (year, month and day, as more precise information are not always available) and separated by a distance shorter than a chosen threshold, are consider as duplicates, except if they belong to the same cruise. In that case, distance between two successive casts may be very small. It is assumed that data sources taken individually are exempt of duplicates.

[23] Results from the duplicate removal process are summarized in Table 2: out of 601,341 casts, 270,949 duplicates were found, mainly between World Ocean Database 2005 (WOD05 hereinafter) and HydroBase2. This result was predictable as HydroBase2 is made up of the 2001 version of the World Ocean Database [*Conkright et al.*, 2002] and several other hydrographic sources.

[24] Working with a combination of sources instead of a unique one lead to an increase of 8 percent in the number of profiles with respect to the WOD05. Changes induced by using this new data set will be addressed in Section 5.3.

### 2.2.2. Vertical Coordinate System

[25] Each database has a vertical coordinate system that is expressed in depth (*m*) or in pressure (*dbar*). In the surface layers, working with any of the two makes nearly no difference, but at greater depths a unique system has to be decided to ensure coherence between measurements.

[26] Elaborated formulas require the knowledge of both temperature and salinity to convert pressure to depth.

However, salinity measurements are less available than temperature. This is why we used a simpler method described by *Saunders* [1981]. This formulation brings correction around 10 m at 1 km depth and around 100 m at 5 km depth. This correction yield results where the error is always smaller than 1 m if the pressure does not exceed 7500 dbar.

### 2.2.3. Quality Control

[27] Even if a priori controlled data are used, further quality control (QC) procedures have to be applied. A first simple QC based on the range of the observations was implemented in order to remove obvious outliers. The ranges for temperature and salinity were obtained for the North Atlantic basin and the Mediterranean Sea following the tables of *Johnson et al.* [2006]. The adaptation of the permitted ranges in the Mediterranean Sea is decisive to allow for higher salinities through the water column. This first step only discarded a negligible amount of the data, but was essential to clean out evident outliers or remaining exclusion values. The second step consisted in checking the vertical gradient of temperature and salinity: observations with gradient values outside a prescribed range were discarded, similar to *Johnson et al.* [2006]. These two steps were performed on the profiles themselves.

[28] The third QC was performed after the vertical interpolation (Section 2.2.4): the domain was decomposed into 5° by 5° tiles in which mean value and standard deviation (STD) were computed. Values falling outside $\pm N_{std}$ standard deviations around the mean were discarded. *Locarnini et al.* [2006] used different values for $N_{std}$, depending whether they were considering coastal or open ocean zone. Typically, $N_{std}$ was assigned a higher value in the tiles containing coasts in order to take into account the higher variability in coastal areas, due for instance to upwelling or river discharge. Nevertheless, we decide to keep the value $N_{std} = 3$ for each tile, as the number of removed data was relatively small: 0.7 percent of the data were considered as outlier using this standard deviation criterion.

### 2.2.4. Vertical Interpolation

[29] The weighted parabolic interpolation (WPI) [*Reiniger and Ross*, 1968] is applied to the profiles to obtain the temperature and salinity fields at standard levels. This method is used in numerous oceanographic climatologies, such as the World Ocean Atlas [*Locarnini et al.*, 2006] and the MEDATLAS [*Fichaut et al.*, 2003]. It is preferred to other methods because it does not generate nonphysical oscillations in the interpolated profiles. The original WPI routine was slightly modified, so that if any value is recorded within 5 m of the surface, this value is directly used as the surface value.

## 2.3. Statistics and Distribution

[30] The distribution of data by decade is shown in Figure 3a: most of the data (72.3%) were acquired after 1970, with a peak between 1980 and 1990 (28.1% of the total number of profiles). The negative trend of the number of available data in the last decades is due to the time lag between data acquisition and their public availability in databases. This is also observed in other similar works [e.g., *Lozier et al.*, 1995; *Gouretski and Koltermann*, 2004].

[31] The monthly distribution (Figure 3b) underlines the contrast between summer and winter regarding the number of available data: August is the month with the most profiles (more than 40703) while December has about 60% less data.
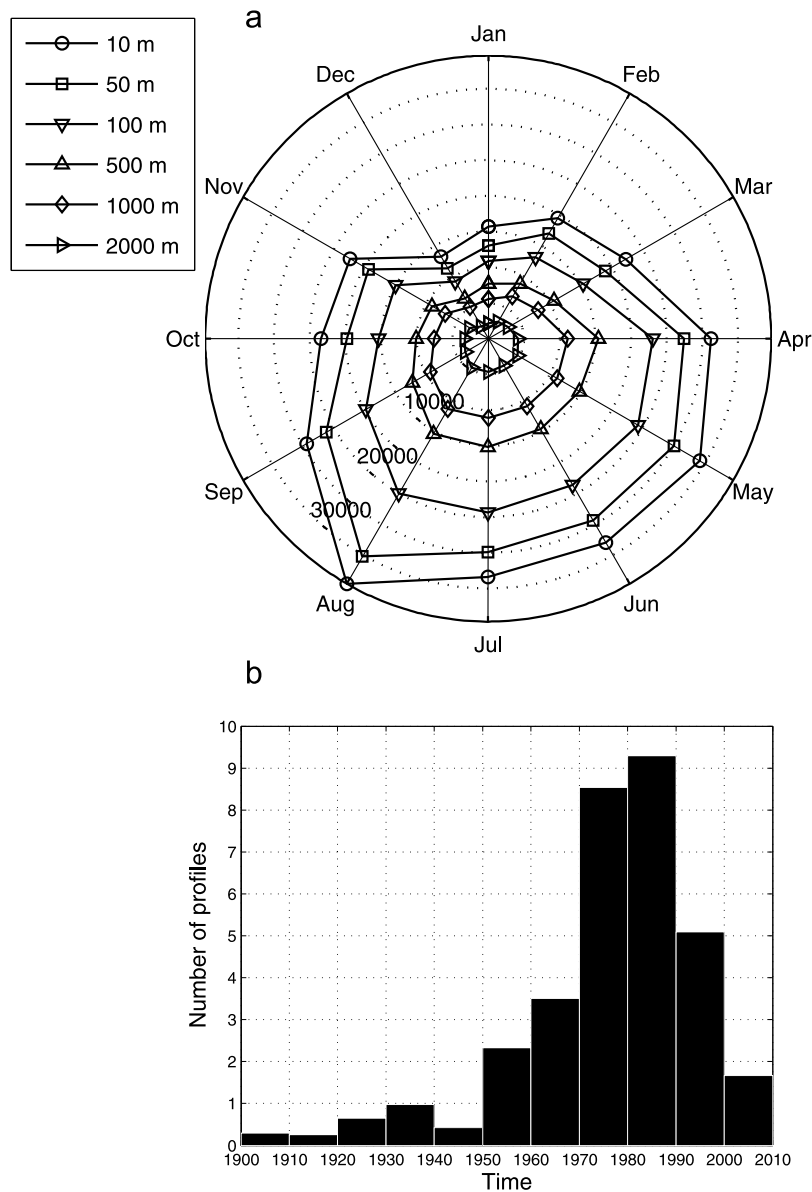
**Figure 3.** (a) Time histogram of data and (b) monthly distribution at various depths; dotted circle are separated by 5000 units.

Similar observations are made on every depth layer. This irregular distribution has to be taken into account when computing annual fields, otherwise results would be biased by the highest data coverage during summer months. For this reason, annual fields are computed by averaging the monthly means. Doing so, the uneven seasonal distribution is partially overcome. On average, 42.6% of the profiles go below 500 m, 31.9% below 1000 m and only 12.9% below 2000 m.

## 3. Method

[32] The code of Diva method is freely available for download at http://modb.oce.ulg.ac.be/viewsvn/, which contains the svn repository (or can be obtained from C. Troupin).

### 3.1. Formulation

[33] Let us consider a set of $N_d$ observations $d_j$, located at positions $\mathbf{r}_j$. The idea is to find a field $\varphi$ sufficiently close to the data, but whose variations are not too large. In a mathematical form, the problem consists of finding $\varphi$ that minimizes the variational principle

$$J[\varphi] = \sum_{j=1}^{Nd} \mu_j \left[ d_j - \varphi(\mathbf{r}_j) \right]^2 + \|\varphi\|^2, \tag{1}$$

with

$$\|\varphi\|^2 = \int_D \left( \alpha_2 \boldsymbol{\nabla}\boldsymbol{\nabla}\varphi : \boldsymbol{\nabla}\boldsymbol{\nabla}\varphi + \alpha_1 \boldsymbol{\nabla}\varphi \cdot \boldsymbol{\nabla}\varphi + \alpha_0 \varphi^2 \right) dD, \tag{2}$$

where $\nabla$ is the gradient operator and $\nabla\nabla\varphi : \nabla\nabla\varphi$ is the squared Laplacian of the field $\varphi$.

[34] The functional is composed of two penalty terms: the first for the smoothness of the solution, the second for the compatibility of the data. The first term of the right-hand side of (1) (compatibility term) takes into account the distance between observations and reconstructed field, so that coefficients $\mu_j$ penalizes misfits between data and analysis; if the variational principle was only composed of this term, the solution would be the pure interpolation of the data, as the minimum would be obtained when $d_j = \varphi(x_j, y_j)$ for $j = 1$, …, $N_d$. As stated in the introduction, this solution is not adapted to hydrographic data.

[35] The second term, made explicit in (2), is the smoothness term. It constitutes a measure of the field regularity. The expression within the integral is a positive definite operator and remains invariant with respect to any rotation of the reference frame [Brasseur and Haus, 1991]. $\alpha_0$ penalizes the field norm, $\alpha_1$ penalizes gradients and $\alpha_2$ penalizes variability. Without loss of generality $\alpha_2$ is chosen to be equal to 1, since function (1) is homogeneous.

[36] This formulation is referred to as Variational Inverse Method (VIM) [Brasseur and Haus, 1991; Brasseur et al., 1996]. Under some conditions, smoothing splines, utilized in the VIM method, and OA are strictly equivalent [McIntosh, 1990; Bennet, 1992], thus it permits to statistically justify the method.

## 3.2. Parameter Determination

[37] Parameters $\alpha_0$, $\alpha_1$ and $\mu_j$ may be determined from the data themselves, following developments from Brasseur et al. [1996]. Let us assume that $\mu_j = \mu$, for $j = 1,…, N_d$, i.e., all data have the same weight.

[38] Writing equations (1) and (2) in nondimensional form (with $\frac{1}{L}\tilde{\nabla} = \nabla$, $L$ being a characteristic length of the problem), we have

$$J[\varphi] = \sum_{j=1}^{Nd} \mu[d_j - \varphi(x_j, y_j)]^2 + \int_{\tilde{D}}\left(\frac{1}{L^4}\tilde{\nabla}\tilde{\nabla}\varphi : \tilde{\nabla}\tilde{\nabla}\varphi + \frac{\alpha_1}{L^2}\tilde{\nabla}\varphi \cdot \tilde{\nabla}\varphi + \alpha_0\varphi^2\right)L^2\,d\tilde{D} \tag{3}$$

and multiplying by $L^2$

$$J[\varphi] = \sum_{j=1}^{Nd} \mu L^2[d_j - \varphi(x_j, y_j)]^2 + \int_{\tilde{D}}\left(\tilde{\nabla}\tilde{\nabla}\varphi : \tilde{\nabla}\tilde{\nabla}\varphi + \alpha_1 L^2\tilde{\nabla}\varphi \cdot \tilde{\nabla}\varphi + \alpha_0 L^4\varphi^2\right)d\tilde{D}. \tag{4}$$

The coefficient $\alpha_0$ fixes the length scale over which variations are significant to move the kernel function of the norm (2) from 1 to 0

$$\alpha_0 L^4 = 1. \tag{5}$$

Here $\mu L^2$ fixes the relative weight on data (signal, $\sigma^2$) versus regularization (noise, $\epsilon^2$)

$$\mu L^2 = 4\pi\frac{\sigma^2}{\epsilon^2} = 4\pi\lambda, \tag{6}$$

where $\lambda$ is the signal-to-noise ratio. The coefficient $\alpha_1$ fixes the influence of gradients

$$\alpha_1 L^2 = 2\xi, \tag{7}$$

where $\xi$ is a nondimensional parameter close to 1 if the gradients are to be penalized with a similar weight than the second derivatives. In the following, $\xi$ is chosen equal to 1. Equations (5) and (6) show that the coefficients of the variational principle are related to the data through two parameters:

[39] 1. The correlation length $L$ translates the distance over which given data influence its neighborhood. It will be determined by fitting the correlation between the data to the theoretical kernel function of (2).

[40] 2. The signal-to-noise ratio $\lambda$ gives an indication of the confidence one can have in a data; thus, it measures the closeness between the data and the analysis. It can be estimated by a generalized cross validation (GCV) technique [Brankart and Brasseur, 1996].

### 3.2.1. Correlation Length

[41] Correlation length is first estimated on each level and for each month by fitting the data correlation function to an analytical form [Brasseur et al., 1996]. Then, for each month, a smoothing filter is applied over the different layers to remove possible irregularities. Such irregularities may arise from data clustering, outliers, or insufficient amount of measurements in a given layer. Finally, the monthly profiles of correlation length are averaged in order to have smooth transition from one month to the other. Indeed, correlation length undergoes large variation from one month to the other. An example is given for the surface: from January to June, estimated correlation length has a value between 1.5 and 1.8°, whereas in September, the reached value is close to 10°. It is believed that such strong variations are due more to the data distribution than to physical reasons, thus the time averaging was applied.

[42] Figure 4 shows the correlation length $L$ as a function of depth: from the surface to around 800–1000 m, $L$ undergoes an increase that may be attributed to the weaker influence of atmosphere and to the presence of more homogeneous water masses. Below 1000 m, $L$ decreases with depth and reaches a minimum at 3500 m. Values for temperature and salinity are similar for most of the water column, as they start to diverge from 3500 to 5500 m depth.

[43] It seems surprising that the value of $L$ at 3500 m is lower than at surface. Nevertheless, we believe that the evolution of $L$ is related to the distributions of water masses, as detailed in Section 4.1. To ensure that it is not an artifact of the data distribution, the same fitting tool was applied to the fields extracted from a high-resolution solution, generated with the ROMS model [Mason, 2009] in the same region of interest. This comparison showed that the variation of $L$ follows a similar evolution in both cases.

[44] The justification of the variability of the water masses distribution with depth would require a more detailed study, but is not in the scope of the present work. Nevertheless, it is assumed that the Mid Atlantic Ridge may also play a role in the separation of the water properties.

[45] Finally, it has to be mentioned that the tool for estimating the correlation length is more efficient when numerous data are available, hence the values computed at
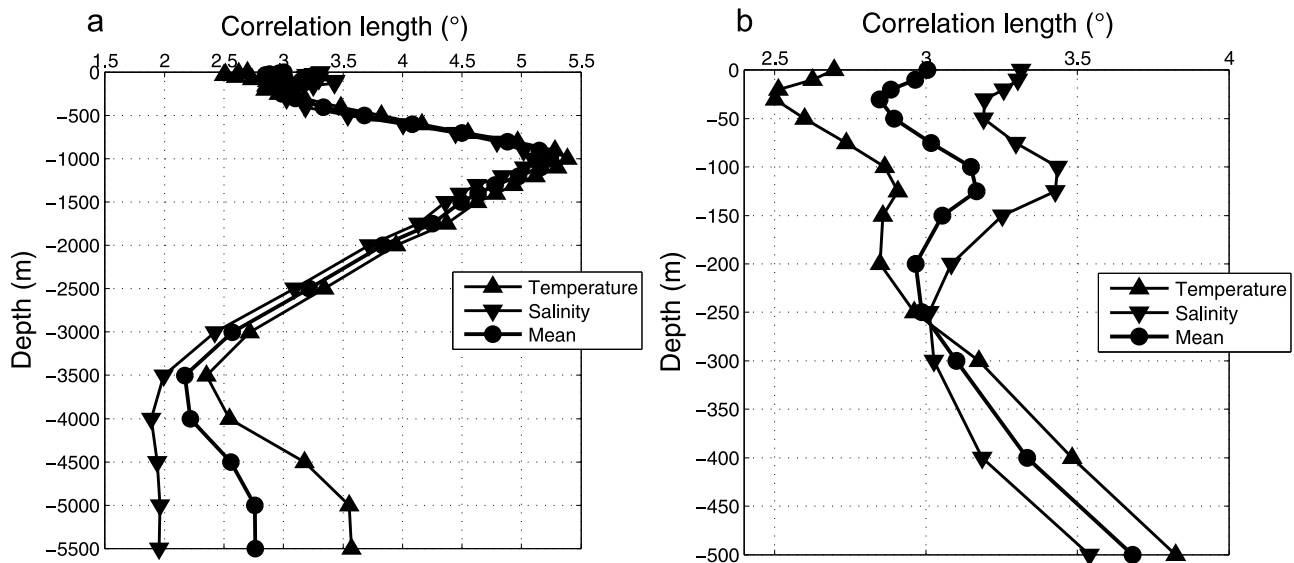
**Figure 4.** Monthly averaged profile of correlation length for temperature and salinity.

higher depths may be affected by a higher degree of uncertainty.

### 3.2.2. Signal-to-Noise Ratio

[46] Analysis were performed with constant and uniform value for signal-to-noise ratio $\lambda = 1$. The values of $\lambda$ provided by cross validation (CV) and Generalized Cross Validation (GCV) were generally too high, so that the resulting analysis fields were too noisy and do not represent climatological conditions, since they display numerous small-scale features or sometimes even reveal the tracks of particular cruises. The choice of $\lambda$ is difficult: in the case of a climatology, the error is not only instrumental, but it also comes from the error of representativity: a climatological field (hence representative of large time scales) is represented using measurements coming from several campaigns (hence having a smaller scale variability). Unfortunately, this kind of error cannot be easily quantified.

[47] The reason why high signal-to-noise ratio values arise from the use of (G)CV is the following: the GCV tries to find the value of $\lambda$ that makes minimal a global estimate of the analysis error variance, $\Theta$, called the generalized cross validator, and defined as follows [*Wahba and Wendelberger*, 1980]:

$$\Theta^2 = \frac{1}{N} \sum_{i=1}^{N} \hat{\theta}_i^2, \qquad (8)$$

where $N$ is the number of data and

$$\hat{\theta}_i^2 = \frac{\left(d_i - \tilde{d}_i\right)^2}{\left(1 - A_{ii}\right)^2}. \qquad (9)$$

In equation (9), $d_i$ is the data at location $\mathbf{x} = \mathbf{x_i}$, $\tilde{d}_i$ is the analysis at $\mathbf{x_i}$ when data $i$ is discarded, and $\mathbf{A}$ is the matrix that links analysis and data, according to

$$\tilde{\mathbf{d}} = \mathbf{A}\mathbf{d}. \qquad (10)$$

[48] The problem of this formulation when applied to oceanographic data, is that such data are not independent, for instance observations of a given cruise made during a relatively short time interval. In that case, GCV yields to very high values for the signal-to-noise ratio, which are incompatible with the nature of a climatology. For this reason $\lambda$ was assigned a constant value.

### 3.3. Background Field

[49] The field $\varphi(\mathbf{r})$ to reconstruct is decomposed into two terms: a background field $\varphi_b(\mathbf{r})$ and a perturbation $\varphi'(\mathbf{r})$. It is the perturbation that is substituted in (2), otherwise the mean value of the field would be also minimized. Hence the background field acts as the solution obtained in regions far away from data. It constitutes a first guess of the field to reconstruct.

[50] In the present implementation, $\varphi_b(\mathbf{r})$ is chosen as a least squares linear regression of the data. Such a background field aims to remove the mean and the spatial trend (e.g. a meridional temperature gradient) from the original data. Configurations using seminorm analysis ($\alpha_0 = 0$) [*Brasseur et al.*, 1996] did not yield significant difference, except near the Amazon River plume: this zone is characterized by strong salinity variations, going from nearly 0 to 30 over a few hundreds kilometers in the first 20 m of the water column [*Lentz and Limeburner*, 1995]. In this particular situation, seminormed background field may lead to unrealistic negative values.

### 3.4. Kernel Function

[51] The Euler-Lagrange differential equation corresponding to the variational principle (1) reads [e.g., *Brasseur et al.*, 1996]

$$\alpha_2 \Delta^2 \varphi - \alpha_1 \Delta \varphi + \alpha_0 = \sum_{j=1}^{N_d} \mu_i (\varphi - d_j) \delta(\mathbf{r} - \mathbf{r}_j), \qquad (11)$$
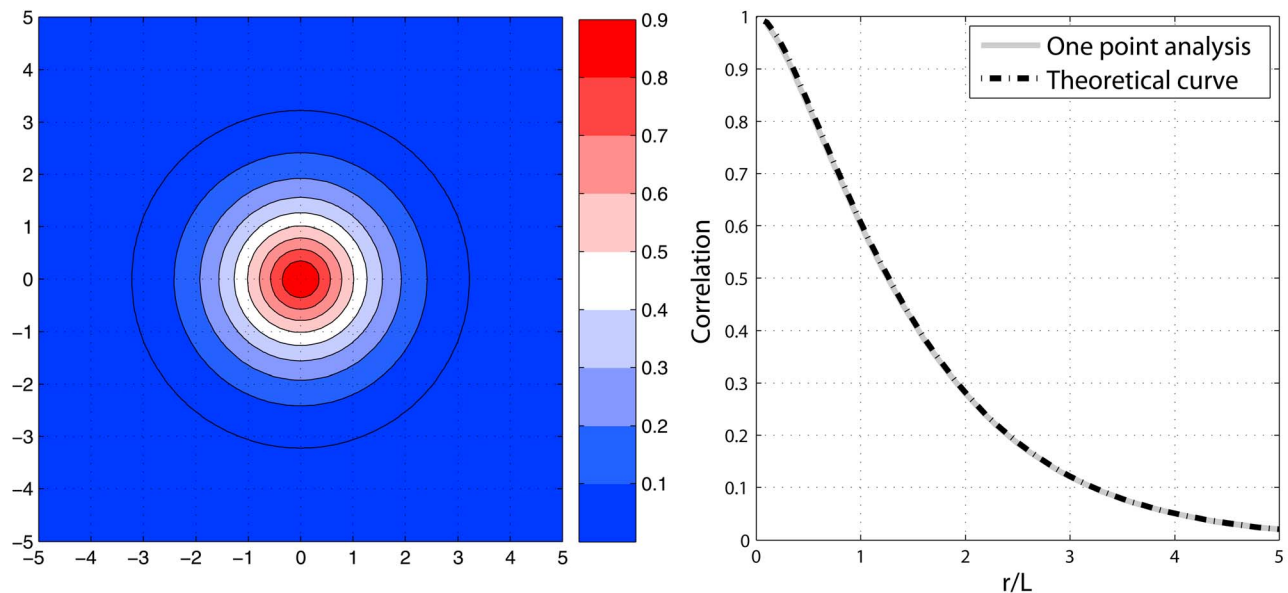
**Figure 5.** (a) Analysis of a single data point with $\lambda = 1000.0$ and no background field. (b) Analysis of a single point along the $x$ axis (gray curve) and the theoretical kernel function (dot-dashed curve).

where $\Delta$ is the Laplacian operator and $\delta$ is the Dirac delta function. If $\mathbf{d}$ is the vector containing the data, the solution of (11) is written as

$$\varphi = \mathbf{k}^T \mathbf{D}^{-1} \mathbf{d},$$

with

$$D_{ij} = K(\mathbf{r}_i, \mathbf{r}_j) + \frac{1}{\mu} \delta_{ij}, \qquad (12)$$

$$k_i = K(\mathbf{r}, \mathbf{r}_i), \qquad (13)$$

where $\delta_{ij}$ stands for the delta of Kronecker ($\delta_{ij} = 1$ if $i = j$, $\delta_{ij} = 0$ if $i \neq j$), $K(\mathbf{r}, \mathbf{r}_i)$ is the kernel of the equation

$$\alpha_2 \Delta^2 \varphi - \alpha_1 \Delta \varphi + \alpha_0 = \delta(\mathbf{r} - \mathbf{r}_j).$$

Let us consider $\alpha_2 = 1$, $\alpha_1 = 2/L^2$ and $\alpha_0 = 1/L^4$ and an infinite domain (where infinite means that boundaries are distant enough from the data in comparison to the correlation length): in this particular case, the solution reads

$$K(\mathbf{r}, \mathbf{r}_j) = K(\mathbf{r} - \mathbf{r}_j) = K(r) = \frac{L^2}{4\pi} \frac{r}{L} K_1\left(\frac{r}{L}\right). \qquad (14)$$

$K_1$ is the modified Bessel function of the second kind [*Abramowitz and Stegun*, 1964, p. 359]. As the domain is infinite, the solution does only depends on the distances $r$ between positions $\mathbf{r}$ and $\mathbf{r}_j$, not on the positions themselves.

[52] For practical purposes, the solution $K$ is not computed, but it is still instructive to represent it in the case of an infinite domain and a unique data with a value of 1, located at the origin. In the first example (Figure 5a), the correlation length is set to 1.0 (arbitrary units): in absence of bound-

aries, the solution is isotropic and goes from 1 at the origin to zero away from the origin.

[53] The form of the kernel is function of the analysis parameter $L$ and is represented in Figure 5b, where $K$ is plotted as a function of the ration $r/L$; the signal-to-noise ratio is assigned a high value ($\lambda = 1000$).

[54] This trivial example illustrates the role of the parameters $L$ and $\lambda$ on the analysis: (1) An increase of $L$ extends the influence region of the data, i.e. the effect of the data is felt at larger distances. (2) Here $\lambda$ does not modify the kernel, but makes the analysis not so close to the observation: there is less confidence in the data, hence the definition of approximation.

### 3.5. Solver

[55] Minimization of (1) is performed by a finite element (FE) method, hence the need for generating a finite element mesh. Since the functional (1) is only defined in the sea, the minimization is limited to the zone defined by the coastline at surface or by the isobath corresponding to the depth of interest, so the FE mesh should only cover that zone (e.g. Figure 6). For practical purposes, the real domain is split into $N_e$ triangular finite elements, so that the variational principle reads

$$J[\varphi] = \sum_{e=1}^{N_e} J_e(\varphi_e). \qquad (15)$$

In each element the solution is a combination of shape functions (3rd order polynomials) and the continuity between elements is assured by identification of adjacent connectors

$$\varphi_e(\mathbf{r_e}) = \mathbf{q_e}^T \mathbf{s}(\mathbf{r_e}), \qquad (16)$$
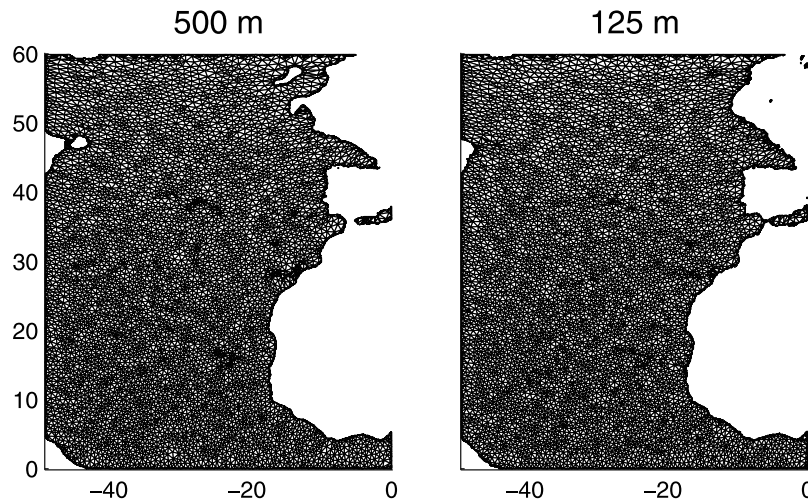
## 500 m            125 m



**Figure 6.** Finite element meshes created at 500 and 125 m. Isobaths are generated from Digital Bathymetric Data Base Variable Resolution (DBDB-V) topography.

with $\mathbf{q}$, the connectors (the new unknowns), $\mathbf{s}$, the shape functions, and $\mathbf{r_e}$, the position in a local coordinate system. Substituting (16) in (15) and using the variational principle (1), it comes

$$J_e(\mathbf{q_e}) = \mathbf{q_e}^T \mathbf{K_e} \mathbf{q_e} - 2\mathbf{q_e}^T \mathbf{g_e} + \sum_{i=1}^{N_{d_e}} \mu_i d_i, \qquad (17)$$

where $\mathbf{K_e}$ is the local stiffness matrix, constructed from the integration of the shape functions, and $\mathbf{g}$ is a vector which depends on local data. On the whole domain, (17) reads

$$J(\mathbf{q}) = \mathbf{q}^T \mathbf{K} \mathbf{q} - 2\mathbf{q_e}^T \mathbf{g_e} + \sum_{i=1}^{N_d} \mu_i d_i, \qquad (18)$$

of which the minimum is reached when

$$\mathbf{q} = \mathbf{K}^{-1} \mathbf{g}. \qquad (19)$$

Matrix $\mathbf{K}$ has a size approximatively proportional to the number of degrees of freedom of the system, but can be very sparse if the elements are properly sorted. In that case the number of operations to invert $\mathbf{K}$ is approximatively proportional to the power 5/2 of the number of degrees of freedom.

[56] To map the data on the finite element mesh, a transfer operator $\mathbf{T_2}$, depending on the shape functions, is applied

$$\mathbf{g} = \mathbf{T_2}(\mathbf{r})\mathbf{d}$$

and finally, to have the solution at any location inside the domain, another transfer $\mathbf{T_1}$ is applied

$$\varphi(\mathbf{r}) = \mathbf{T_1}(\mathbf{r})\mathbf{q}.$$

Combining the two previous equations, the relation between $\varphi$, the interpolated field at location $\mathbf{r}$, and the data vector $\mathbf{d}$ is obtained:

$$\varphi = \mathbf{T_1}(\mathbf{r})\mathbf{K}^{-1}\mathbf{T_2}(\mathbf{r})\mathbf{d}. \qquad (20)$$

[57] The asset of the resolution by FE is twofold: (1) from the point of view of physics: data can only influence a zone that is physically connected to it through the mesh; (2) from the point of view of numerics: for a large number of data, the method requires less operations than OA.

[58] It is instructive to compare the meshes in Figure 6 (left) and Figure 6 (right): at 500 m, Mediterranean Sea and Atlantic Ocean are separated, whereas at 125 m, the finite element mesh link one zone to the other. Implications of this property are analyzed in Section 5.1.

### 3.5.1. Topography and Coastlines

[59] Coastlines are necessary to delimit the area covered by the FE mesh. Reference depths are chosen from surface to 5500 m, similar to other large-scale climatologies. Depth contours are created using the Digital Bathymetric Data Base Variable Resolution (DBDBV, US Navy), a 5 min resolution bathymetry. General Bathymetric Chart of the Oceans (GEBCO) 1 min topography was also considered, but for the region of sake it generated too many small contours, which make the FE mesh difficult to generate. Use of high-resolution bathymetries is more advisable for less extended regions.

### 3.5.2. Finite Element Mesh

[60] Once the contours are created, meshes are constructed on each of the 33 levels (Figure 6) with a scale $L_e = 1/3°$. This value is a compromise between resolution and computational time, since mesh generation is a costly operation. As within each triangular element, the solution is a combination of shape functions (third order polynomials), features with scale approximatively larger or equal to $L_e/3$ can be resolved. Meshes are generated only once and then stored for further operations.

### 3.6. Outlier Detection

[61] An automatic outlier detection is implemented within Diva. It is based on a comparison between the data analysis residual and the expected standard deviation (J.-M. Beckers et al., Advanced data interpolating variational analysis: Application to a hydrographic climatology, manuscript in preparation, 2009). This outlier detection algorithm has the

asset of being applied during the analysis itself, so it does not require a priori information on the data. It will be showed in Section 5 that this step is essential to guarantee the quality of the analysis.

## 4. Results

[62] Temperature and salinity fields generated with the methodology described in Section 3 and data from Section 2 have been produced on monthly, seasonal and annual basis, on 33 depth levels. Gridded fields of 0.1° resolution are displayed on the GHER server at http://gher-diva.phys.ulg. ac.be/web-vis/clim.html in the directory GHER, with the name NE Atlantic and are available for download at http://gher-diva.phys.ulg.ac.be:8080/GHER/NEAtlantic/ (OPeNDAP) and at http://gher-diva.phys.ulg.ac.be/data/GHER/NEAtlantic/.

### 4.1. Annual Fields

[63] The annual fields are computed by averaging all the monthly fields. Several levels are mapped in Figure 7 to illustrate some of the outputs. An exhaustive description of water properties and formation across the entire domain is not in the objective of the present paper. The interested reader will find this information in other relevant publications [e.g., *Lozier et al.*, 1995; *van Aken*, 2000a, 2000b, 2001]. The objective is limited to a more simple description of typical temperature and salinity fields, with a focus on the link between property dispersion diagram and data-derived parameters presented in Section 3.2.

[64] To have a first validation of the results, the *T-S* diagrams presented here were compared to those obtained from the WOA05 and from a high-resolution numerical model covering the same region [*Mason*, 2009]. In the three cases, the diagrams display comparable features concerning the water mass distribution. Then the observations made in the next sections about the water mass structure seem quite coherent.

#### 4.1.1. Near-Surface Fields

[65] The highest salinity values are observed in the central part of the subtropical gyre and in the Mediterranean Sea, where evaporation overcomes freshwater flux. Four zones are characterized by low salinity: 1. coastal zones of France and England, 2. the southwest part the domain, 3. the Gulf of Guinea and 4. the Labrador region. The three formers undergo the effects or river discharge, while the latter has combined effects of melting ice and rivers.

[66] The general temperature structure is a meridional gradient due to different solar radiation. Exception are in Labrador Sea and in the northwest Africa region, where upwelling is visible from Gibraltar Strait to Cape Blanc. In the northwestern part of the domain, strong temperature and salinity gradients appear, and a front separates cold, fresh waters and warm, salty waters. These strong gradients remain visible up to about 700 m.

[67] Another front is observed south of the subtropical gyre (Figure 7), where the westward North Equatorial Current (NAC) makes the connection between South Atlantic Water and North Atlantic Central Water [*Lozier et al.*, 1995].

[68] *T-S* diagram (Figure 8 (left)) evidences the presence of numerous water masses as well as the high surface variability, consequence of the air-sea exchanges. This may

constitute an explanation for the low values observed for the correlation lengths with respect to higher depth values (Figure 4). The strong scattering of properties is mainly observed in the first 100 m of the water column.

#### 4.1.2. Intermediate Depth Fields

[69] At 700 m, a region of warmer water is limited to the south by a front between North Atlantic and South Atlantic waters and to the north by strong gradients described in the previous section. The salinity map from Figures 7f and 7h reveals a clear separation between high-salinity Mediterranean Waters (MW) and Atlantic waters. Between 1000 and 1500 m, the maximal values of temperature and salinity appear near the Iberian Peninsula and Gibraltar.

[70] The lower scattering of properties is underlined by *T-S* diagram (Figure 8 (middle)), where a limited number of water masses seems to be present at this level. Such situation is representative of depth levels between 200 and 1750 m, with an overall tendency of decreasing dispersion when depth is increased. As stated before, the same observations were made on the *T-S* diagrams made using WOA and high-resolution model fields (not shown here).

#### 4.1.3. Deep Water Fields

[71] At higher depths, variations of temperature and salinity are very weak with respect to what is observed at surface. Physical separation between western and eastern Atlantic basins is evidenced by the field differences across the boundary, even if small amounts of North East Atlantic Deep Water (NEADW) is able to flow though the small gaps of the mid-Atlantic ridge.

[72] In the NW part of the domain, a zone of low temperature (Figure 7i) is assumed to be the signature of Denmark Strait Overflow Water (DSOW), which crosses Denmark Strait (between Greenland and Iceland) to reach the Labrador Sea [*Lazier et al.*, 2002]. Highest temperatures are located in the northern zone of the eastern basin, centered west of Spain, and is attributed to North Atlantic Deep Water (NADW).

[73] Figure 8 (right) shows broad scattering in the property distribution, in agreement with the correlation length plot (Figure 4) where the minimal value appears around 3500 both for temperature and salinity. Similar *T-S* diagrams are found between 2500 and 5500 m.

### 4.2. Error Fields

[74] Along with the analyzed field, it is often instructive to have the associated error fields, reflecting the confidence one can have in the results. Basically, the error is expected to depend on two factors: the data coverage and the noise on data. OA method provides a way to compute an error estimate, according to the formula

$$e^2(\mathbf{r}) = \sigma^2 - \mathbf{g}(\mathbf{r})^T \mathbf{D}^{-1} \mathbf{g}(\mathbf{r}), \qquad (21)$$

where $\sigma^2$ is the variance of the true anomaly field, $\mathbf{g}(\mathbf{r})$ is the covariance of the data with respect to the true field (which is dependent of the location $\mathbf{r}$), and $\mathbf{D}$ is the data covariance matrix. For large data sets, such as the ones used for climatology production, the operation (21) is costly, since it requires the inversion of the covariance matrix $\mathbf{D}$.

[75] *McIntosh* [1990] and *Bennet* [1992] demonstrated the equivalence between smoothing splines and OA, under the condition of identity of covariance function of OA and re-
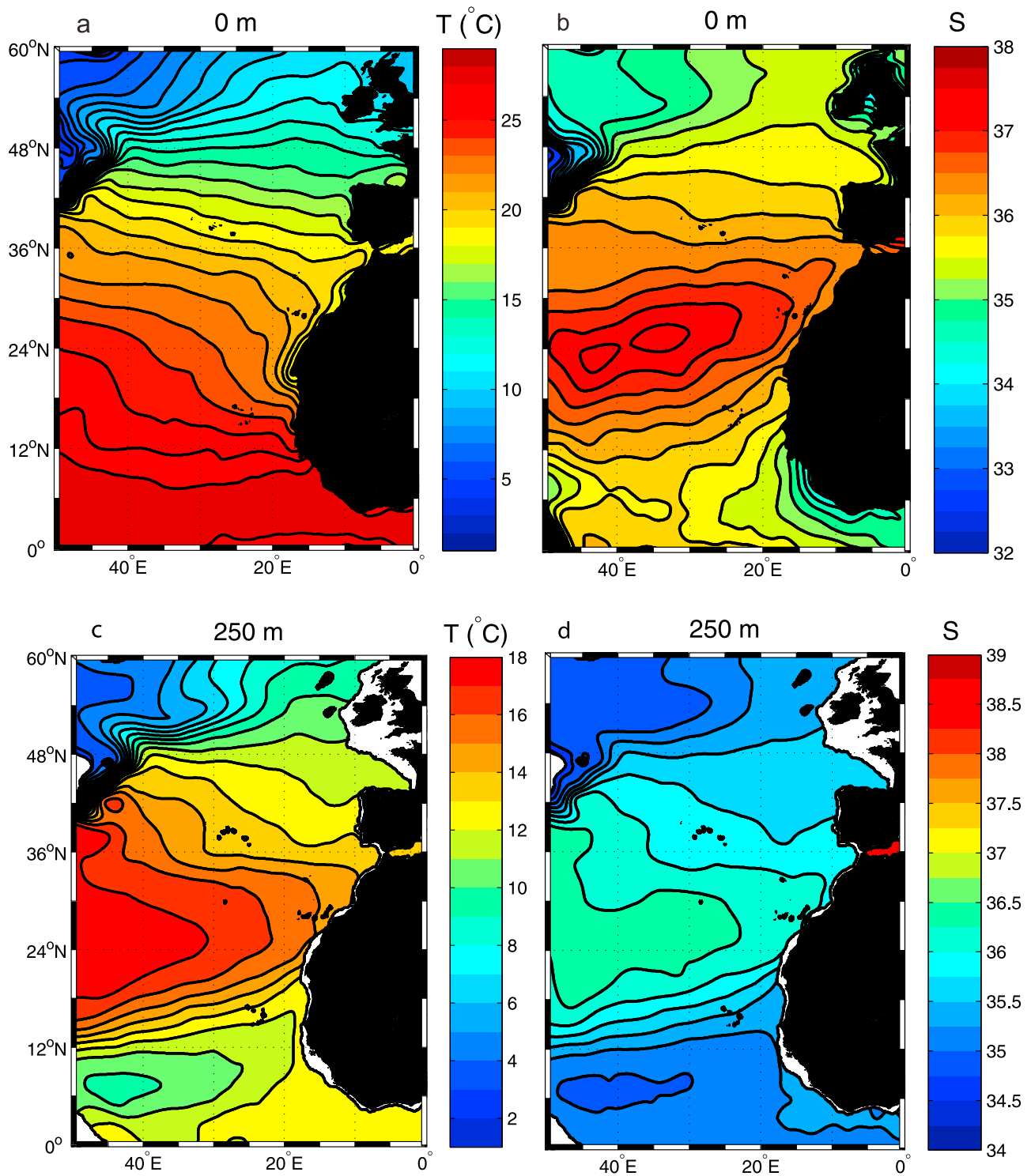
**Figure 7.** Annual temperature (left) and salinity fields (right) at 0, 250, 700, 1300, and 3500 m. Black surfaces indicate surface land mask while white surfaces represent depth contours. Isotherms are separated by 1°C, except at 3500 m where $\Delta T = 0.1$°C. Isohalines are separated by 0.25 in all cases, except at 3500 m where $\Delta S = 0.05$.
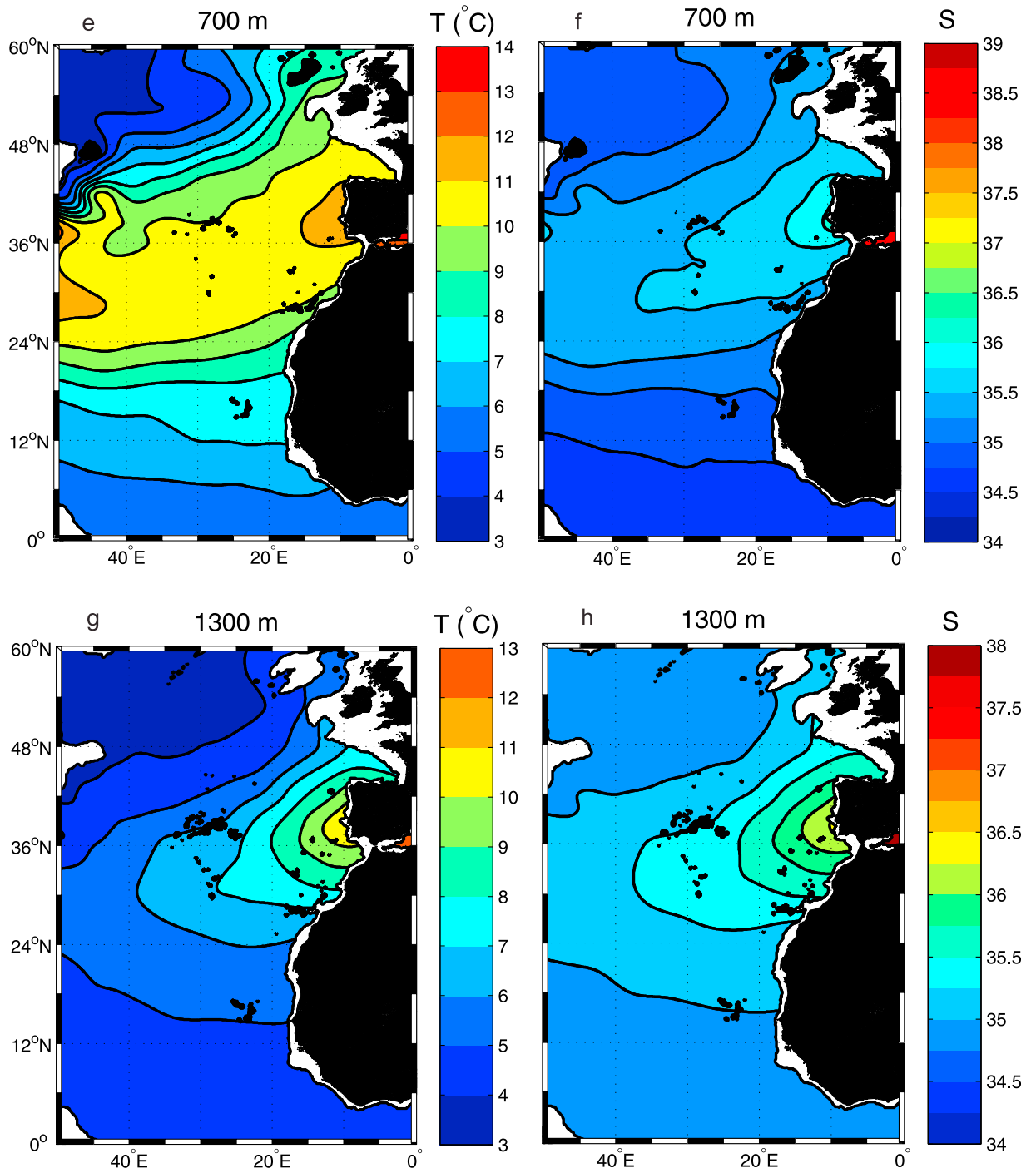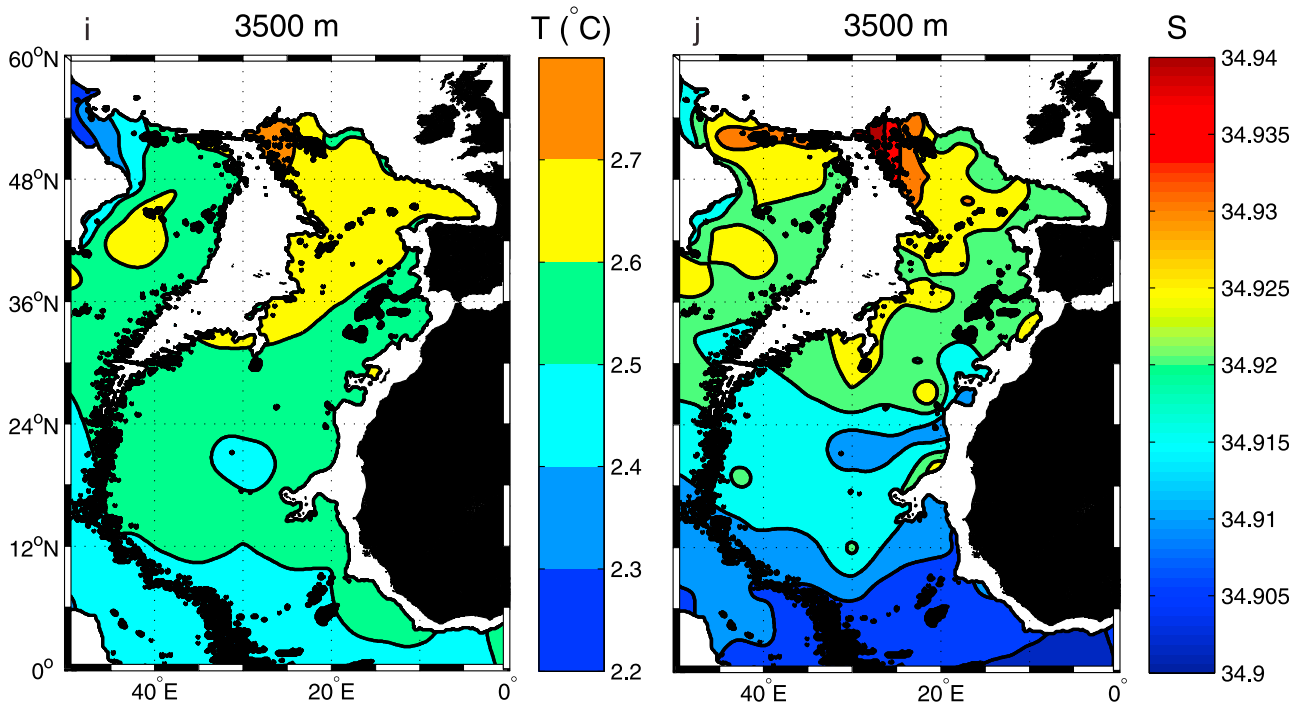
**Figure 7.** (continued)

**Figure 7.** (continued)

producing kernel for norm splines. *Brankart and Brasseur* [1998] and *Rixen et al.* [2000] used this identity to derive an error expression for the VIM. In practice the data input of the analysis tool for an error calculation is a vector containing the covariance of data points with respect to the point where the error estimate has to be calculated. *Rixen et al.* [2000] showed that, for large grids, VIM becomes numerically cheaper than OA when the number of data is larger than the number of degrees of freedom of the finite element mesh. This method will be referred to as hybrid method and is one of the three implemented in Diva.

[76] In OA, $\mathbf{g(r)}^T\mathbf{D}^{-1}$ is the analysis operator, i.e. the operator that relates the data ($\mathbf{d}$) to the analyzed field. To compute the error field, the analysis operator has to be applied to $\mathbf{g(r)}$ (21). The problem comes down to evaluate $\mathbf{g(r)}$.

[77] Recent developments from Beckers et al. (manuscript in preparation, 2009) permitted the derivation of the real covariance function through the execution of Diva. At each point at which the error is needed, two analyzes are performed: the first one is made with the already existing **LU** decomposition (decomposition into a product of a lower triangular matrix **L** and an upper triangular matrix **U**) of one Diva execution to calculate the error, the other with an existing **LU** decomposition of another Diva execution to calculate the covariance function.

[78] An additional simplification provides the poor man's error indicator: instead of applying the analysis operator to $\mathbf{g(r)}$, it is applied to a vector of the size of the data, but filled with $\sigma^2$. It constitutes the quickest method, since the error is evaluated at every grid points in only one operation. Nevertheless, the error fields computed this way are systematically underestimated.

[79] An application is presented in Figure 9: relative (to the field variance) error fields for temperature are computed on a $0.5 \times 0.5°$ grid for the temperature in January at 2000 m. In Figure 9, the standard deviation (in °C) is represented.
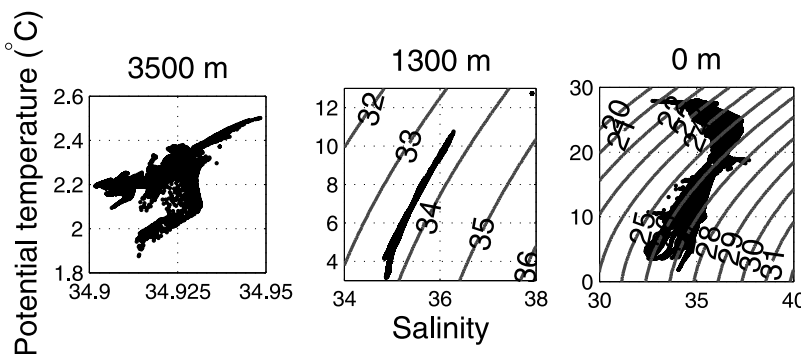


**Figure 8.** *T-S* diagram at (right) 0, (middle) 1300, and (left) 3500 m.
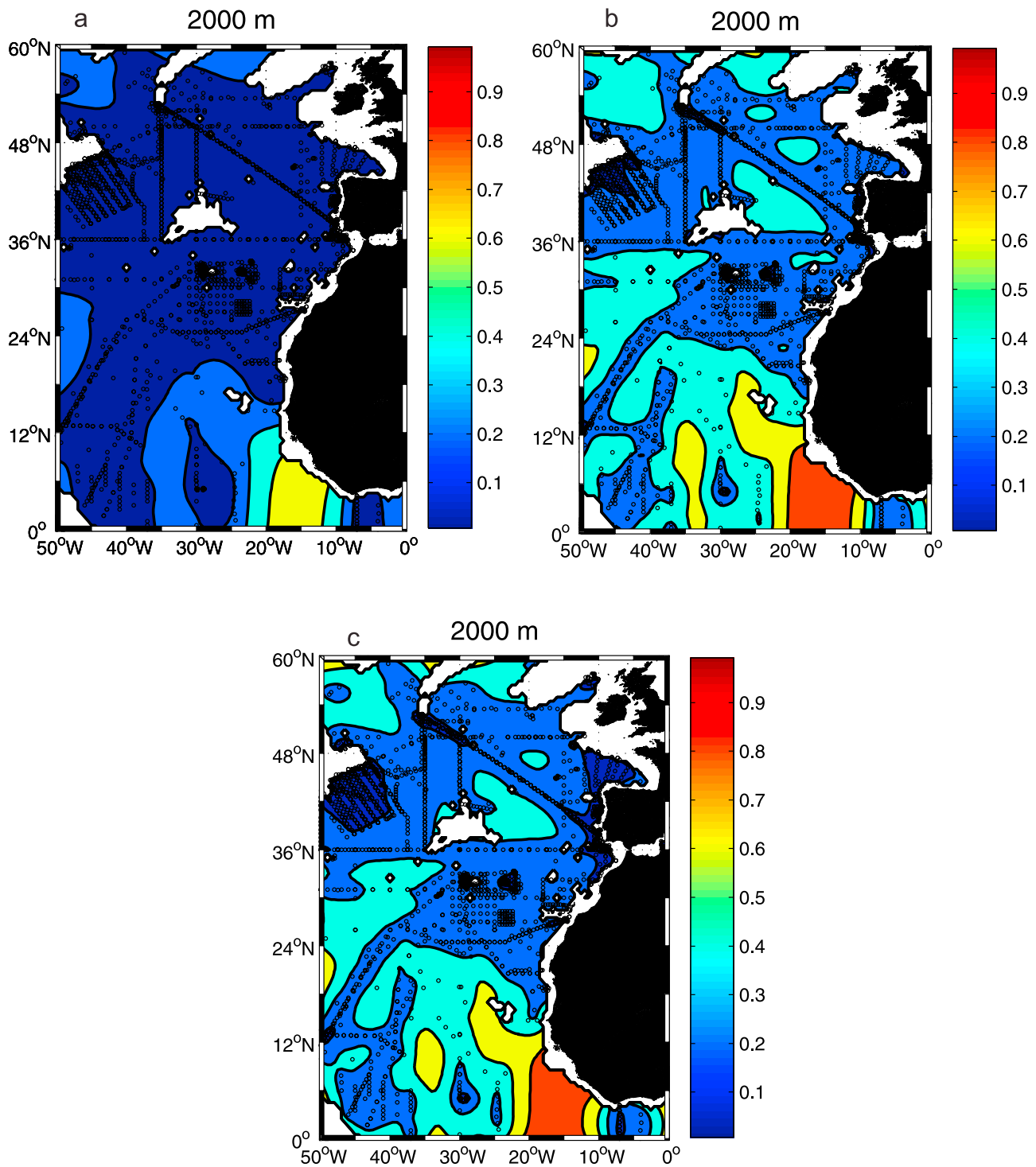
**Figure 9.** Relative error fields computed with the different methods for the temperature in January at 2000 m: (a) poor man's error estimate, (b) hybrid method, and (c) real covariance function. Black circles indicate data locations. Isolines are distant of 0.2.

In each case, the importance of data coverage is underlined, as the main error region occur in a zone that approximatively spreads from 10 to 20°W and from the equator to 12°W and totally void of data. Another location where high errors are observed is within the small isolated contours within the Middle Atlantic Ridge. The difference with the previous region is that these small contours are not physi-

cally connected to the rest of the ocean in the horizontal layer, illustrating the disconnection of error field in nonadjacent areas.

[80] Figures 9b and 9c show very similar error fields, meaning that in the present case, the hybrid approach was sufficient to correctly estimate the covariance function. The poor man's version yields lower values over the whole
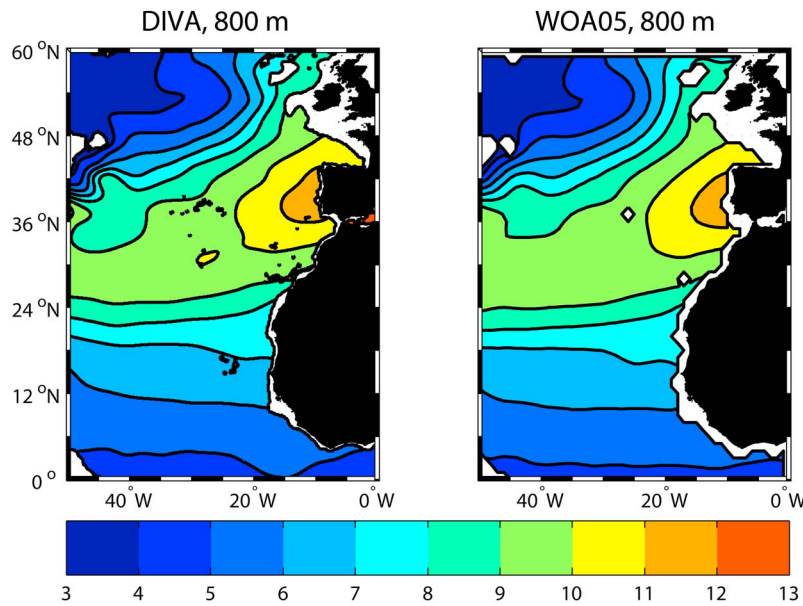
**Figure 10.** Annual temperature fields at 800 m (left) from Diva and (right) from WOA05.

domain. However, it efficiently provides a qualitative mapping of the error estimate, which can later be used to produce a mask over the analyzed fields.

[81] Eventually, similar error computations were carried out at different levels and at different times of the year, both for temperature and salinity. They revealed the same conclusions concerning the error distribution and the importance of data coverage.

## 5. Discussion

### 5.1. Comparison With World Ocean Atlas

[82] To have an overall validation of our method, our maps are compared with the 1° resolution WOA05 (Figure 10). Even if this first comparison is mainly visual, it confirms that without sophisticated quality control or any application of posterior filtering on the fields, Diva is able to reproduce features very similar to the WOA05.

[83] However, the lack of resolution of WOA05 (and other recent climatologies, such HydroBase and WOCE

GHC) is particularly visible near the coasts and around archipelagos (e.g., Canary Islands, Azores). Also, interior seas and channels (Mediterranean Sea, English Channel, Strait of Gibraltar) are not covered at all by the latter climatologies.

### 5.1.1. Separation Between Basins

[84] Salinity analysis around the Strait of Gibraltar are presented in Figure 11. The objective of this example is to illustrate the mixing of waters with OA (WOA05) and Diva methods. In the right panel (WOA05), a zonal gradient visible on both side of the strait: the radius of influence of the OA method has propagated the high-salinity information westward, while the Mediterranean waters are supposed to sink and flow into the Atlantic Ocean at higher depths. The analysis provided by Diva is different, as the flow of less saline Atlantic water into the Mediterranean Sea is clearly visible.

[85] Another instructive example is given by *Brasseur et al.* [1996], where in situ measurements in the Mediterranean Sea are analyzed using both OA and VIM. In the OA
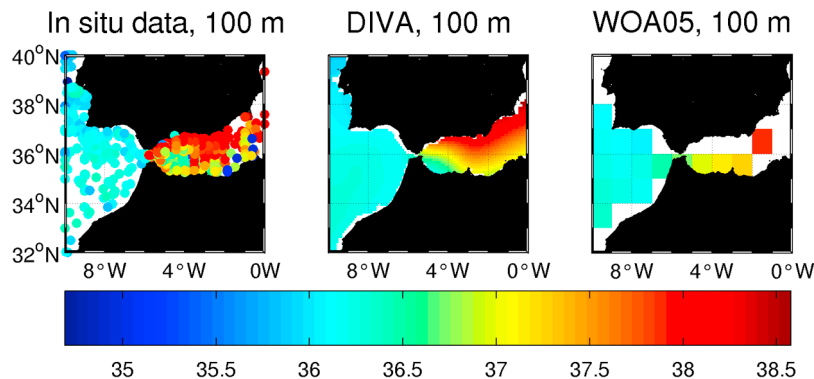


**Figure 11.** Salinity observations in July at 100 m around the Strait of Gibraltar and corresponding analysis from Diva climatology and WOA05.
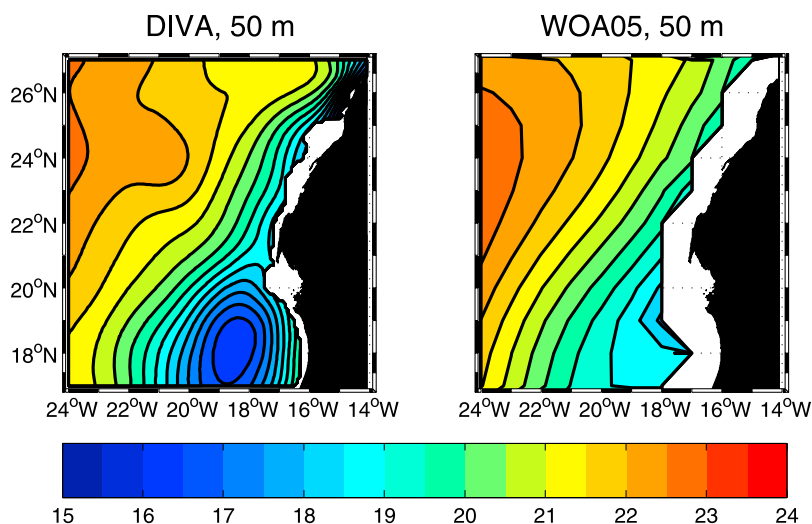
**Figure 12.** Temperature fields in August at 50 m depth (a) from Diva climatology and (b) from WOA05.

solution, it appears that low-salinity coastal waters in the Adriatic Sea influences waters in the Tyrrhenian Sea, while the two are separated by Italy.

### 5.1.2. Coastal Area

[86] This particular example concentrates on Cape Blanc (20°N, Mauritania) upwelling region, in order to show the impact of a better resolution near the coast. It is essential to stress that this illustration did not require a new analysis, but is merely an enlargement of a determinate zone, in order to show features that would not visible on the global map.

[87] The Cape Blanc region is characterized by a permanent upwelling [e.g., *Barton et al.*, 1998] and a filament [*Van Camp et al.*, 1991; *Gabric et al.*, 1993]. This structure is known to be quasi-permanent and has been frequently sampled, making available a large number of data. South of the Cape (between 15°N and 21°N) lies a mesoscale cyclonic eddy. Its position varies over the year, but it is still present all year long.

[88] To illustrate these features, the temperature field is extracted at 50 m in August from the Diva and the WOA05 climatologies (Figure 12). North of the Cape, the characteristics of the fields are similar, except that in the left panel, the temperature reaches lower values, characteristic of the coastal upwelling. However, this first difference can be, at least partly, attributed to the grid resolution of the two climatologies, and to the mode of representation of the fields (contours).

[89] A more interesting feature is the gyre visible south of the Cape in the left panel. As the data sets used are not too different (Section 5.3), the difference between the two fields is attributed to the analysis method itself: the WOA05 uses 1° square temperature averages as input of the objective analysis program [*Locarnini et al.*, 2006]. The averaging is responsible for the smoothing of the field and may generate the loss of information contained in the data.

### 5.2. Outlier Detection

[90] As part of the analysis, an automatic quality control routine is applied. The objective is to remove data that create irregularities in the analyzed fields but that were not previously detected by a priori quality checks. Such data may not be defined as outliers, since their value is situated within an acceptable range and since they do not differ from the local (5° × 5°) mean with more than three STD's. But they are not compatible with the construction of a climatology, which supposes relatively regular fields, i.e., exempt of very local variation.

[91] An application is shown in Figure 13: on the right side (no QC), several small salinity structures (indicated by black arrows), referred as bull's eye by *Lozier et al.* [1995], are visible, for example south of the Azores. These structures may be attributed to suspect data and the latter shall be removed in the final version of the climatological field, even if the measurements are correct. On the left is the field after the removal of the suspect observations: the bull's eyes have been eliminated.

[92] Typically, these observations may arise from campaigns that aimed to sample small scale structures (eddies, meddies, filaments etc). If the number of such observations is small with respect to the total number of data in the region, Diva quality control will be able to detect them automatically. Otherwise, if these observations represent a significant proportion of the available data in the region, they will be considered as good data, thus the small scale structure will be present in the final analyzed field. In this case, data have to be removed by hand. A concrete example is a meddy sampled during a cruise in June 1982, with a clear signature from 700 to 1300 m depth.

[93] The difference between the two fields (with and without outliers) is mapped in Figure 14, along with the suspect data locations. The first striking fact is that most of the suspect data are located near Newfoundland and Labrador shelves. As outlined in Section 2.1, this zone is characterized by a strong variability, due to the encounter of Gulf Stream and Labrador Current. Nevertheless, the removal of numerous data in this region did not strongly affect the analyzed field. This is simply explained by the large number of measurements available (see Figure 2).

[94] A second important feature is that the greatest differences between the two fields occur in the southwestern part
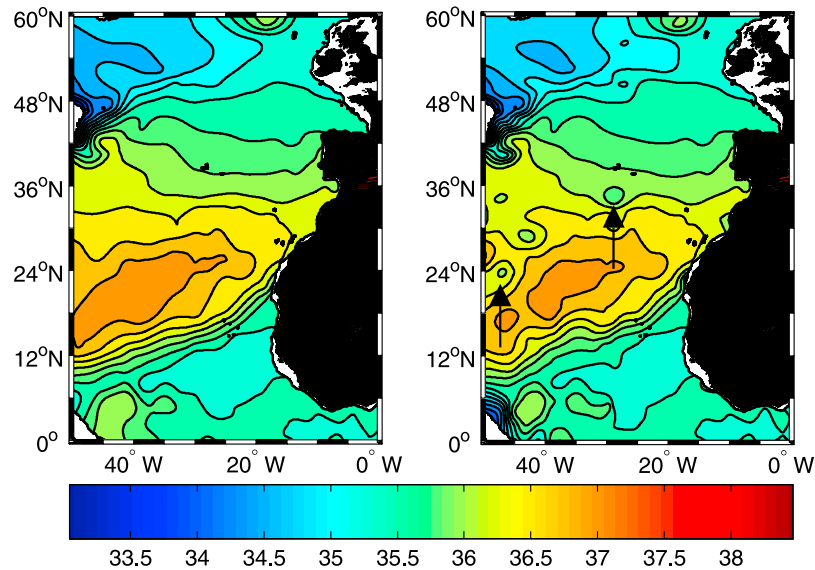
**Figure 13.** Salinity field in January at 125 m depth (left) with and (right) without Diva quality check.

of the domain, next to the Brazilian shelf. The color bar reveals that waters are saltier in the case where outliers are removed. This is attributed to the localized influence of the Amazon fresher water, transported northwestward by the North Brazil Current.

[95] It is to highlight that the data responsible for the small salinity patches were detected, even if their number was relatively important in the case of the south Azores region. This section is concluded by the important comment that the decision of removing or not suspect data has to be made by the user himself. The decision should be made in agreement with the pursued goal, considering that analyzing climatological data is different than working with synoptic data sets. Generally, a good knowledge of the physical processes in the region of interest is a necessary step toward the finalization of a climatology.

### 5.3. Benefits From Using Aggregated Data Set

[96] As described in Section 2, the climatology presented in this paper is constructed upon an aggregated data set, including the WOD05. Two questions arise from this specificity: (1) Is it worth working with such data sets, or put in other words, is it worth dedicating time to produce new data sets from several sources? (2) How does it translate, in terms of temperature and salinity, to the analyzed fields? The answers to these questions obviously depend on the location of the additional data with respect to the original database, WOD05 in this case, and on their quality.

[97] Analyzed fields are compared with and without the additional observations, as illustrated in Figure 15, which shows temperature fields in October at 700 m. These fields are representative of general features that were obtained for other depths, months or variables. Two possibilities have to be considered: (1) additional data are found where data coverage is already good and (2) additional data manage to fill in the gaps left by the initial coverage.

[98] For the domain of interest, it is mainly the situation described in the first case: the initial data coverage was

already good, and the additional observations were nearly always located in already sampled zones. This is why the two fields are relatively close (lower than 0.1°C) within a large area. Nevertheless, large differences between the fields occur locally, for instance south of the Grand Banks of Newfoundland (between 36°N and 42°N). Justifications to these differences are quite delicate. A first hypothesis is that
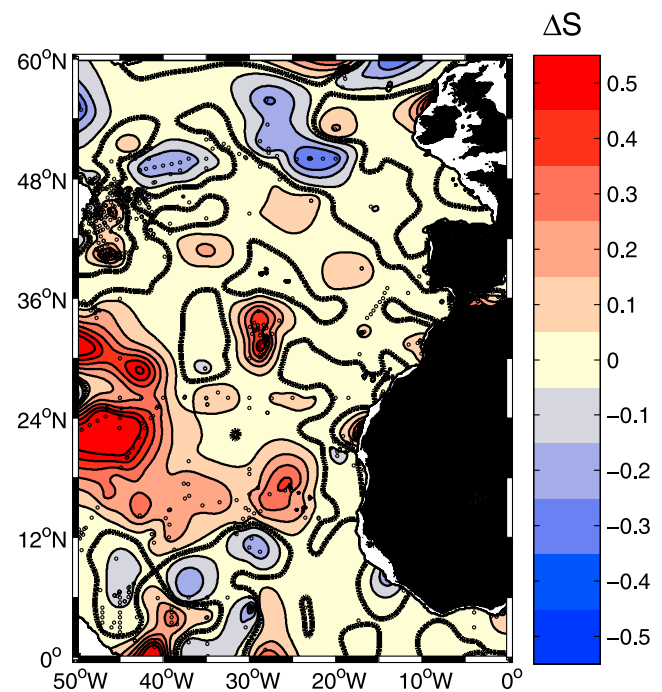


**Figure 14.** Salinity difference between analyzed fields without and with suspect data. Suspect observations are denoted by black circles. The isohalines are separated by $\Delta S = 0.1$, thick lines show the contours where $\Delta S = 0$. Red color indicates higher salinity for the no outlier field.
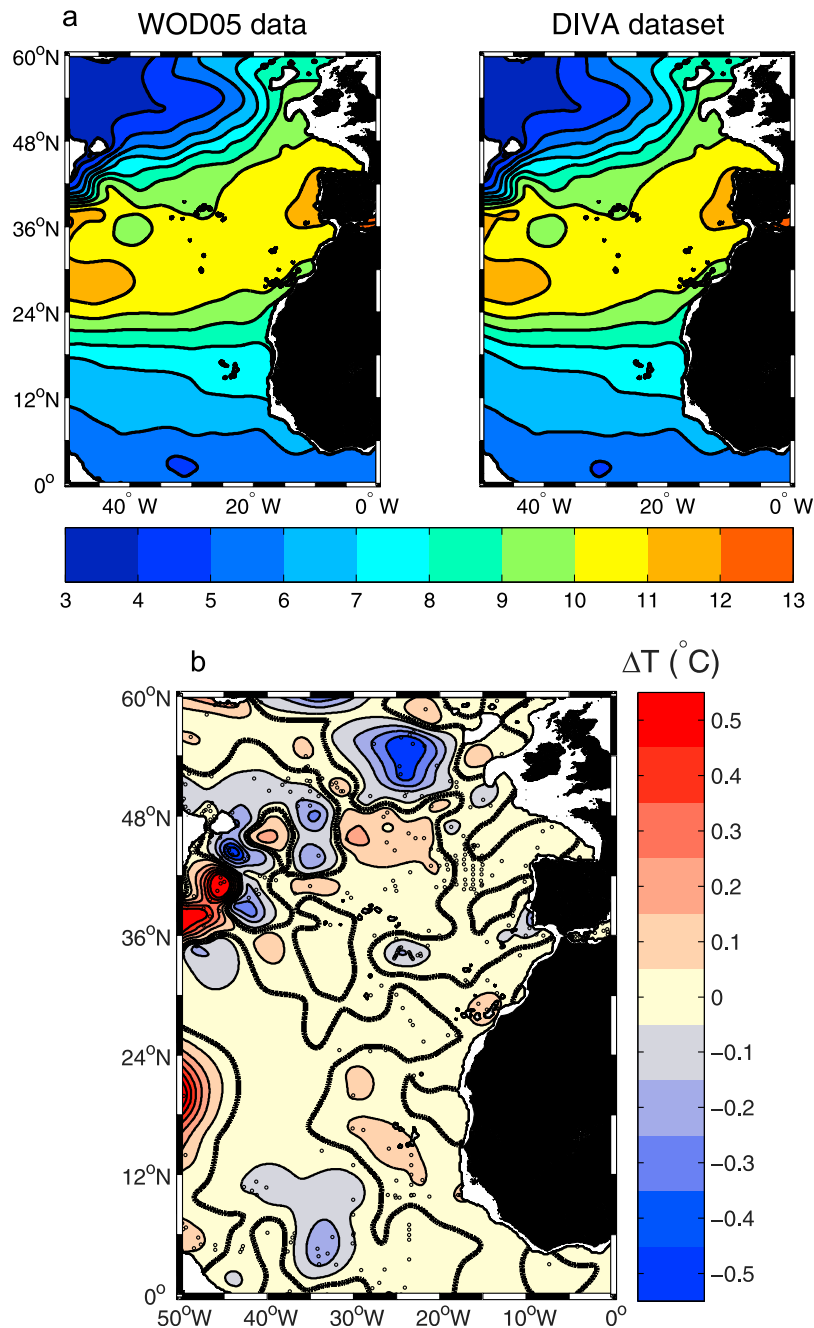
**Figure 15.** (a) Temperature fields in October at 700 m depth, as obtained from (left) WOD and (right) merged data sets (isotherms are separated by 1°C). (b) Difference between the two fields (isotherms separated by $\Delta T = 0.1$°C, thick lines show the contours where $\Delta T = 0$).

data have undergone stricter quality control before their inclusion into WOD than for the other individual databases (Section 2.1). The second hypothesis is that these data are indeed correct and have to be included in the analysis. For example, if the measurements were made during campaigns under particular oceanographic or atmospheric conditions, they might perturb the field constructed using observations from the initial data set. As the Newfoundland region is known for its strong variability, the latter hypothesis turns out to be plausible. The solution to confirm one of the hypothesis would be to apply again the WOD quality checks to all the other databases prior to their merging.

## 6. Conclusions

[99] We present a practical application of the software Diva on an aggregation of several data sets in the northeast Atlantic. Diva possesses numerous advantages over recently

developed climatologies: this objective method is applicable in any region of the ocean and implies a limited number of parameters needed to go from raw data to the gridded climatology. The innovation in the present work is twofold: 1. the database, constructed by assembling data sets from various sources and eliminating the resulting duplicates; 2. the method, of which the benefits were reviewed along this paper: high resolution output grid, consideration of physical boundaries in the analysis process, low number of parameters, determined in objective way, efficient finite element solver, and generation of associated error fields. The latter is of particular importance, as most of currently available climatologies do not offer this feature.

[100] Comparisons with WOA05 reveal a better representation of the coastal zones, whereas they are usually the most sampled regions. Also, the physical separation of water masses (i.e., through bottom topography) showed in Section 5.1 highlights the coherence of the variational method.

[101] Concerning the relevance of using aggregated database, the results obtained here did not show significant improvements, that is to say the overall data coverage was only slightly enhanced, since the new data were mostly situated in regions already well sampled. However, this small amount of extra data was sufficient to create visible differences in the analyzed fields. It remains to evidence if this is due to a real addition of information from the new data, or if it is an effect of insufficient quality control in each individual database. Nevertheless, the construction of local climatologies on more reduced regions may certainly benefit from the use of gathered data sets.

[102] Recently, numerical experiments using ROMS model [*Shchepetkin and McWilliams*, 2005] in the eastern subtropical Atlantic, yielded encouraging results, namely improvements in the representation of the Azores Current when Diva climatology is used for initialization and extraction of boundary condition, instead of WOA05 (E. Mason, personal communication, 2009). Also, the correct detection of outliers was of primary importance to assure physically acceptable fields.

[103] Finally, the method outlined here is very general, hence it can be applied to any region of the World Ocean, provided the data have been properly prepared.

[104] Future work will consist in the extension of the selected types of instruments for the new database: they will not be limited only to CTD's and bottles, but will also include profiling floats, XBT's and MBT's. These constitute valuable sources of information on the ocean characteristics and are expected to help increasing the data coverage, thus decreasing the error fields. Once again, the obstacle to overcome is the preparation of data, for instance the quality control and correction of instrument systematic errors. Nevertheless, the first tests performed to create this new database showed that the additional observations are mostly concentrated where data are already present. This is why it is believed that the general results will not undergo substantial modifications.

[105] Another interesting direction to follow is the use of isopycnic coordinates instead of isobaths. The principle of Diva is to perform analysis in horizontal planes, whatever the vertical coordinates are. Then the task will consist in transforming depths in density prior to the analysis and then to switch back to depths after it.

## References

Abramowitz, M., and I. A. Stegun (1964), *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, Dover, New York.

Barton, E. D., et al. (1998), The transition zone of the Canary Current upwelling region, *Prog. Oceanogr.*, *41*, 455–504, doi:10.1016/S0079-6611(98)00023-8.

Bennet, A. (1992), *Inverse Methods in Physical Oceanography*, 368 pp., Cambridge Univ. Press, Cambridge, U. K.

Boyer, T., S. Levitus, H. Garcia, R. A. Locarnini, C. Stephens, and J. Antonov (2005), Objective analyses of annual, seasonal, and monthly temperature and salinity for the World Ocean on a 0.25° grid, *Int. J. Clim.*, *25*, 931–945, doi:10.1002/joc.1173.

Brankart, J.-M., and P. Brasseur (1996), Optimal analysis of in situ data in the western Mediterranean using statistics and cross-validation, *J. Atmos. Oceanic Technol.*, *13*, 477–491, doi:10.1175/1520-0426(1996)013<0477:OAOISD>2.0.CO;2.

Brankart, J.-M., and P. Brasseur (1998), The general circulation in the Mediterranean Sea: A climatological approach, *J. Mar. Syst.*, *18*, 41–70, doi:10.1016/S0924-7963(98)00005-0.

Brasseur, P. (1994), Reconstruction de champs d'observations océanographiques par le Modèle Variationnel Inverse: Méthodologie et applications, Ph.D. thesis, Univ. of Liège, Liège, Belgium.

Brasseur, P., and J. Haus (1991), Application of a 3-D variational inverse model to the analysis of ecohydrodynamic data in the northern Bering and southern Chukchi seas, *J. Mar. Syst.*, *1*, 383–401, doi:10.1016/0924-7963(91)90006-G.

Brasseur, P., J.-M. Beckers, J.-M. Brankart, and R. Schoenauen (1996), Seasonal temperature and salinity fields in the Mediterranean Sea: Climatological analyses of a historical data set, *Deep Sea Res. Part I*, *43*, 159–192, doi:10.1016/0967-0637(96)00012-X.

Calder, C. A., and N. Cressie (2007), Some topics in convolution-based spatial modeling, paper presented at 56th Session of the International Statistics Institute, Inst. Nac. de Estat., Lisbon.

Chang, C.-W. J., and Y. Chao (2000), A comparison between the World Ocean Atlas and Hydrobase climatology, *Geophys. Res. Lett.*, *27*, 1191–1194, doi:10.1029/1999GL002379.

Chassignet, E. P., L. T. Smith, G. R. Halliwell, and R. Bleck (2003), North Atlantic simulations with the Hybrid Coordinate Ocean Model (HYCOM): Impact of the vertical coordinate choice, reference pressure, and thermobaricity, *J. Phys. Oceanogr.*, *33*, 2504–2526, doi:10.1175/1520-0485(2003)033<2504:NASWTH>2.0.CO;2.

Cleveland, W. S., and S. J. Devlin (1988), Locally weighted regression: An approach to regression analysis by local fitting, *J. Am. Stat. Assoc.*, *83*, 596–610.

Conkright, M., et al. (2002), World Ocean Database 2001 CD-rom data set documentation, technical report, Natl. Oceanogr. Data Cent., Silver Spring. Md.

Fichaut, M., M. Garcia, A. Giorgetti, A. Iona, A. Kuznetsov, M. Rixen, and MEDAR Group (2003), MEDAR/MEDATLAS 2002: A mediterranean and black sea database for operational oceanography, in *Building the European Capacity in Operational Oceanography, Proceedings of the Third International Conference on EuroGOOS*, vol. 69, edited by H. Dahlin et al., pp. 645–648, doi:10.1016/S0422-9894(03)80107-1, Elsevier, Amsterdam.

Gabric, A. J., L. García, L. van Camp, L. Nykjaer, W. Eifler, and W. Schrimpf (1993), Offshore export of shelf production in the Cap Blanc giant filament as derived from CZCS imagery, *J. Geophys. Res.*, *98*, 4697–4712, doi:10.1029/92JC01714.

Gandin, L. S. (1965), *Objective Analysis of Meteorological Fields*, Gidrometeorol. Izd., Leningrad.

Gouretski, V. V., and K. P. Koltermann (2004), WOCE—Global hydrographic climatology: A technical report, *Ber. 35/2004*, 52 pp., Bundessamtes für Seeschiffahrt und Hydrogr., Hamburg, Germany.

Hanawa, K., and T. Yasuda (1992), New detection method for XBT depth error and relationship between depth error and coefficients in the depth-time equation, *J. Oceanogr.*, 48, 221–230, doi:10.1007/BF02239006.

Hanawa, K., P. Rual, R. Bailey, A. Sy, and M. Szabados (1994), Calculation of new depth equations for expendable bathythermographs using a temperature-error-free method (Application to Sippican/TSK T7; T6; and T4 XBTs), *Tech. Ser. 42*, Intergov. Oceanogr. Comm., Paris.

Higdon, D. M. (1998), A process-convolution approach to modelling temperatures in the North Atlantic Ocean, *Environ. Ecol. Stat.*, 5, 173–190, doi:10.1023/A:1009666805688.

Higdon, D. M., J. Swall, and J. Kern (1998), *Bayesian Statistics 6*, Oxford Univ. Press, Oxford, U. K.

Johnson, D. R., T. P. Boyer, H. E. Garcia, R. A. Locarnini, A. V. Mishonov, M. T. Pitcher, O. K. Baranova, J. I. Antonov, and I. V. Smolyar (2006), World Ocean Database 2005 documentation, *Intern. Rep. 18*, 163 pp., Natl. Oceanogr. Data Cent., Silver Spring, Md.

Karafistan, A., J.-M. Martin, M. Rixen, and J.-M. Beckers (2002), Space and time distributions of phosphates in the Mediterranean Sea, *Deep Sea Res. Part I*, 49, 67–82, doi:10.1016/S0967-0637(01)00042-5.

Kim, H.-S., A. Gangopadhyay, L. K. Rosenfeld, and F. L. Bub (2007), Developing a high-resolution climatology for the central California coastal region, *Cont. Shelf Res.*, 27, 2135–2161, doi:10.1016/j.csr.2007.05.011.

Lazier, J., R. Hendry, A. Clarke, I. Yashayaev, and P. Rhines (2002), Convection and restratification in the Labrador Sea, 1990–2000, *Deep Sea Res.*, 49, 1819–1835, doi:10.1016/S0967-0637(02)00064-X.

Lemos, R. T., and B. Sansó (2006), Spatio-temporal variability of ocean temperature in the Portugal Current System, *J. Geophys. Res.*, 111, C04010, doi:10.1029/2005JC003051.

Lemos, R. T., and B. Sansó (2009), A spatio-temporal model for mean, anomaly, and trend fields of North Atlantic sea surface temperature, *J. Am. Stat. Assoc.*, 104(485), 5–18, doi:10.1198/jasa.2009.0018.

Lentz, S. J., and R. Limeburner (1995), The Amazon River Plume during amasseds: Spatial characteristics and salinity variability, *J. Geophys. Res.*, 100(C2), 2355–2375, doi:10.1029/94JC01411.

Levitus, S. (1982), Climatological atlas of the World Ocean, *NOAA Prof. Pap. 13*, U.S. Gov. Print. Off., Washington, D. C.

Levitus, S., and T. P. Boyer (1994), *World Ocean Atlas 1994*, vol. 4, *Temperature*, *NOAA Atlas NESDIS*, vol. 4, 129 pp., NOAA, Silver Spring, Md.

Locarnini, R. A., A. V. Mishonov, J. I. Antonov, T. P. Boyer, and H. E. Garcia (2006), *World Ocean Atlas 2005*, vol. 1, *Temperature*, *NOAA Atlas NESDIS*, vol. 61, edited by S. Levitus, 182 pp., NOAA, Silver Spring, Md.

Lozier, M. S., W. B. Owens, and R. G. Curry (1995), The climatology of the North Atlantic, *Prog. Oceanogr.*, 36, 1–44, doi:10.1016/0079-6611(95)00013-5.

Marchesiello, P., J. C. McWilliams, and A. Shchepetkin (2003), Equilibrium structure and dynamics of the California Current System, *J. Phys. Oceanogr.*, 3, 753–783, doi:10.1175/1520-0485(2003)33<753:ESADOT>2.0.CO;2.

Mason, E. (2009), High-resolution modelling of the Canary Basin oceanic circulation, Ph.D. thesis, Univ. de Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain.

McIntosh, P. C. (1990), Oceanographic data interpolation: Objective analysis and splines, *J. Geophys. Res.*, 95(C8), 13,529–13,541, doi:10.1029/JC095iC08p13529.

Reiniger, R., and C. Ross (1968), A method of interpolation with application to oceanographic data, *Deep Sea Res.*, 15, 185–193, doi:10.1016/0011-7471(68)90040-5.

Ridgway, K. R., J. R. Dunn, and J. L. Wilkin (2002), Ocean interpolation by four-dimensional weighted least squares—Application to the waters around Australasia, *J. Oceanic Atmos. Technol.*, 19, 1357–1375, doi:10.1175/1520-0426(2002)019<1357:OIBFDW>2.0.CO;2.

Rixen, M., J.-M. Beckers, J.-M. Brankart, and P. Brasseur (2000), A numerically efficient data analysis method with error map generation, *Ocean Modell.*, 2, 45–60, doi:10.1016/S1463-5003(00)00009-3.

Saunders, P. M. (1981), Practical conversion of pressure to depth, *J. Phys. Oceanogr.*, 11, 573–574, doi:10.1175/1520-0485(1981)011<0573:PCOPTD>2.0.CO;2.

Shchepetkin, A. F., and J. C. McWilliams (2005), The regional oceanic modeling system (ROMS): A split-explicit, free-surface, topography-following-coordinate oceanic model, *Ocean Modell.*, 9, 347–404, doi:10.1016/j.ocemod.2004.08.002.

Stephens, C., J. I. Antonov, T. P. Boyer, M. E. Conkright, R. A. Locarnini, T. D. O'Brien, and H. E. Garcia (2002), *World Ocean Atlas 2001*, vol. 1, *Temperature*, *NOAA Atlas NESDIS*, vol. 49, NOAA, Silver Spring, Md.

Swall, J. (1999), A process convolution approach to modeling nonstationary spatial depence, Ph.D. thesis, Duke University, Durham, N. C.

van Aken, H. M. (2000a), The hydrography of the mid-latitude northeast Atlantic Ocean: Part I: The deep water masses, *Deep Sea Res. Part I*, 47, 757–788, doi:10.1016/S0967-0637(99)00092-8.

van Aken, H. M. (2000b), The hydrography of the mid-latitude northeast Atlantic Ocean: Part II: The intermediate water masses, *Deep Sea Res. Part I*, 47, 789–824, doi:10.1016/S0967-0637(99)00112-0.

van Aken, H. M. (2001), The hydrography of the mid-latitude northeast Atlantic Ocean: Part III: The subducted thermocline water mass, 48, 237–267, doi:10.1016/S0967-0637(00)00059-5.

Van Camp, L., L. Nykjaer, E. Mittelstaedt, and P. Schlittenhardt (1991), Upwelling and boundary circulation off northwest Africa as depicted by infrared and visible satellite observations, *Prog. Oceanogr.*, 26, 357–402, doi:10.1016/0079-6611(91)90012-B.

Wahba, G., and J. Wendelberger (1980), Some new mathematical methods for variational objective analysis using splines and cross validation, *Mon. Weather Rev.*, 108, 1122–1143, doi:10.1175/1520-0493(1980)108<1122:SNMMFV>2.0.CO;2.

A. Barth, J.-M. Beckers, M. Ouberdous, D. Sirjacobs, and C. Troupin, GeoHydrodynamics and Environment Research, AGO, University of Liège, Allée du 6 Août 17, B-4000 Liège, Belgium. (ctroupin@ulg.ac.be)

F. Machín, Facultad de Ciencias del Mar, Departamento de Física, Universidad de Las Palmas de Gran Canaria, E-35017 Las Palmas de Gran Canaria, Spain.