A SPACE-TIME MODEL FOR JOINT MODELING OF OCEAN TEMPERATURE AND SALINITY LEVELS AS MEASURED BY ARGO FLOATS

SUJIT K. SAHU, PETER CHALLENOR

ABSTRACT

The world's climate is to a large extent driven by the transport of heat and fresh water in the oceans. Regular monitoring, studying, understanding and forecasting of temperature and salinity at different depths of the oceans are a great scientific challenge. Temperature at the ocean surface can be measured from space. However salinity cannot yet be measured by satellites, and space-based measurements can only ever give us values at the surface. Until recently temperature and salinity measurements within the oceans have had to come from expensive research ships. The Argo float program has been funded by various nations to collect actual measurements and rectify this problem.

A Bayesian hierarchical model is proposed in this paper describing the spatio-temporal behaviour of the joint distribution of temperature and salinity levels. The model is obtained as a kernel-convolution effect of a single latent spatio-temporal process. Additional terms in the mean describe non-stationarity arising in time and space. Predictive Bayesian model selection criteria have been used to validate the models using data for the year 2003. Illustrative annual prediction maps along with their uncertainty maps are also obtained. The Markov chain Monte Carlo methods are used throughout in the implementation.

Southampton Statistical Sciences Research Institute
Methodology Working Paper M07/14

A space-time model for joint modeling of ocean temperature and salinity levels as measured by Argo floats

Sujit K. Sahu and Peter Challenor*

The world's climate is to a large extent driven by the transport of heat and fresh water in the oceans. Regular monitoring, studying, understanding and forecasting of temperature and salinity at different depths of the oceans are a great scientific challenge. Temperature at the ocean surface can be measured from space. However salinity cannot yet be measured by satellites, and space-based measurements can only ever give us values at the surface. Until recently temperature and salinity measurements within the oceans have had to come from expensive research ships. The Argo float program has been funded by various nations to collect actual measurements and rectify this problem.

A Bayesian hierarchical model is proposed in this paper describing the spatio-temporal behavior of the joint distribution of temperature and salinity levels. The model is obtained as a kernel-convolution effect of a single latent spatio-temporal process. Additional terms in the mean describe non-stationarity arising in time and space. Predictive Bayesian model selection criteria have been used to validate the models using data for the year 2003. Illustrative annual prediction maps along with their uncertainty maps are also obtained. The Markov chain Monte Carlo methods are used throughout in the implementation.

Key Words: Hierarchical model; Markov chain Monte Carlo; non-stationary spatio-temporal process; North Atlantic; oceanography.

1 Introduction

The Argo float program, see for example www.argo.ucsd.edu, is designed to measure the temperature and salinity of the upper 2,000 meter of the ocean globally. Eventually it is planned to have 3,000 floats deployed across all ocean basins and during 2006 that number should have been deployed. The North Atlantic already has the planned

*Sujit K. Sahu is senior lecturer, School of Mathematics, Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, UK. (Email: S.K.Sahu@soton.ac.uk) and Peter Challenor is senior scientist, National Oceanography Centre, Southampton, UK. (Email: P.Challenor@noc.soton.ac.uk)

number, while more distant oceans such as the South Pacific have until recently been under-represented. Each float is programmed to sink to a depth of 1,000 meters, drifting at that depth for about 10 days. After this period the float sinks a further kilometer to a depth of 2,000 meter and adjusting its buoyancy rises to the surface, measuring temperature and conductivity (from which we derive salinity) on the way. Once at the surface, the data and the position of the float are transmitted via a satellite, this gives scientists access to near real-time data. After transmitting the data the float sinks back to its 'resting' depth of 1,000 meter and drifts for another ten days before measuring another temperature and salinity profile at a different location. Argo data are freely available via the international Argo project office, see the above website. The focus of this paper is modeling and analysis of temperature and salinity data obtained from the Argo floats.

In recent years there has been a tremendous growth in the statistical models and techniques to analyze spatio-temporal data. Such data arise in many contexts e.g. air pollution monitoring, disease mapping, economic data monitoring and so on. Often the primary interests in analyzing such data are to smooth and predict time evolution of some response variables over a certain spatial domain. Typically, such predictions are made from data observed on a large number of variables which themselves vary over time and space. Wikle (2003) provides a review of space time modeling and related issues in environmental science. More recently Gelfand *et al.* (2004) and (2005) describe spatial process modeling for univariate and multivariate spatial data. See also Sahu and Mardia (2005a), Sahu *et al.* (2006, 2007) and the references therein for a recent snapshot of research activities in this area.

The recently developed techniques of space time modeling have not yet been applied to the joint modeling of temperature and salinity levels observed in world oceans. Indeed, we are not aware of any such empirical model describing both spatial and temporal variation in the literature. Higdon (1998) models only temperature data obtained from research vessels (not Argo floats). Ferrari and Polzin (2005) explore the role of different air-sea fluxes which influence the relationships between temperature and salinity at the surface using data from the North Atlantic, see also other research papers co-authored by Ferrari including Ferrari and Rudnick (2000).

The main difficulty in the modeling problem here lies in the fact that any particular location in the ocean is never re-sampled. The Argo data are not typical examples of some number of time series observed in fixed monitoring stations. Thus the statistical techniques often used for analyzing multiple time series data from environmental and other land based sources, see e.g. Sahu and Mardia (2005b), Sahu *et al.* (2007) are no longer useful for modeling Argo data. Analysis methods based on some sort of spatial aggregation, e.g. averaging over some grid-areas, are also unlikely to work here since we may have only a few (much less than 10) data points observed in any particular day (the basic time unit we work with) in the North Atlantic. Aggregation over time may also create unavoidable problems such as heterogeneity. Moreover, such aggregation will completely mis-align the one-to-one relationship between temperature and salinity.

The primary objective of this paper is to build models for high resolution space-time multivariate data on ocean temperature and salinity. Toward this end we adopt a kernel convolution approach initially discussed by Ver Hoef and Barry (1998). In a

series of papers Higdon and his co-authors has popularized the approach by considering a discrete version of the approach, see for example Higdon (1988). We extend his approach in several ways as we jointly model daily temperature and salinity data from the Argo floats.

The remainder of this paper is organized as follows. Section 2 presents some important exploratory analyses of the data in order to facilitate model development. The proposed model is developed in Section 3. Bayesian prediction methods and development of trend analysis are detailed in Section 4. Model based analyses are provided in Section 5. A brief summary and future issues to explore are given in Section 6. An appendix contains the joint and conditional posterior distributions needed for computations.

2 Data description

Let $Z_1(\mathbf{s}, t)$ and $Z_2(\mathbf{s}, t)$ denote the temperature and salinity levels observed at location $\mathbf{s}$ and at time t. Assume that there are N such pairs of observations. The time points at which data are observed are not equi-lagged and we do not assume this in our modeling endeavor. Moreover, it is also possible that all the N locations where data have been observed can be different because of the moving floats. Thus every observation is associated with a particular location $\mathbf{s}$ identified by the latitude (s_1) and longitude (s_2) pair and a time t which is a particular day. For convenience, we shall use $p = 1, \ldots, N$ to index the N observations so that a particular value of p is associated with a particular value of $\mathbf{s}$ for the spatial locations and a particular value of t for the time points.

We consider the data observed in the North Atlantic ocean between the latitudes 20^o and 60^o north and 10^o and 50^o west. We model all valid data observed in the year 2003. Three data sets are created for three different layers of the ocean.

Our first data set consists of 2374 data points observed in the top layer of depth less than 50 meters. We call this the surface data set. Our second data set of 2726 observations is composed of all the observations in between the depths of 475 and 525 meters. This is the mid-layer data. The third data set consists of 2628 observations in between the depths of 975 and 1025 meters. We set aside 250 randomly chosen observations from each of the three data sets for validation purposes. Exploratory analysis has shown that there is only negligible variation in the data due to differences of depths upto 50 meters, as a result we ignore such variation in our subsequent modeling.

2.1 Data at the surface

The spatial locations of the $N = 2374$ observations at the surface are plotted as points in Figure 1. Note that there is exactly one recording of observation at each of the location. The scatter plot of the temperature against salinity levels is given in the top panel of Figure 2. This sort of quadratic relationship between temperature and salinity is well known in the literature, see for example Ferrari and Rudnick (2000). There are, however, a few outlying points which can arise for various reasons including data taken from near the mouth of a river (low salinity), and possible errors in data collection. The

first row of Figure 3 shows monthly seasonal variability in temperature and salinity. The temperature levels vary more than the salinity levels from month to month.

2.2 Data at the mid-layer

The spatial locations for the $N = 2726$ observations at the mid-layer are different from those at the surface. However, the location plot had many similar characteristics as Figure 1 and is omitted for brevity. The scatter plot of the temperature against salinity levels is given in the middle panel of Figure 2. This plot again shows the same type of quadratic relationship between temperature and salinity as seen in the top panel for the ocean surface. However, as expected the variability in both temperature and salinity has decreased greatly, see also the range of the X-axis in each plot. The points concentrate mainly near a theoretical quadratic relationship and there are only a few possible outliers. The second row of Figure 3 shows almost negligible seasonal effects in temperature and salinity as expected. In our modeling we do not include the seasonal terms.

2.3 Data at the deep-layer

As in the mid-layer we have omitted the plot showing the spatial locations of the $N = 2628$ observations at the deep-layer. The scatter plot of the temperature against salinity levels is given in the bottom panel of Figure 2. This plot again shows the same type of quadratic relationship between temperature and salinity as seen previously for the surface and mid-layer. There were not much seasonal variability either in temperature or in salinity. The average temperature and salinity levels, however, in the months of January and February were slightly lower than the same for the other months. As in the mid-layer there would not be much seasonal variation at the deep ocean and we do not include the seasonal terms in our modeling approach.

3 Models

We first assume the hierarchical structure:

$$Z_j(\mathbf{s}, t) = Y_j(\mathbf{s}, t) + \epsilon_j(\mathbf{s}, t), \quad j = 1, 2, \tag{1}$$

where $Y_j(\mathbf{s}, t), j = 1, 2$ are space-time processes (described below) and the error terms $\epsilon_j(\mathbf{s}, t)$ are independent white noise processes assumed to follow $N(0, \sigma_j^2)$. Each space-time process $Y_j(\mathbf{s}, t)$ is modeled as the sum of a mean process, $\mu_j(\mathbf{s}, t)$, and a spatio-temporal process, $v_j(\mathbf{s}, t)$, i.e.

$$Y_j(\mathbf{s}, t) = \mu_j(\mathbf{s}, t) + v_j(\mathbf{s}, t), \quad j = 1, 2. \tag{2}$$

These processes are described below.

3.1 Modeling the mean process

Ocean temperature and salinity are both affected by several factors including the latitude and longitude of the location where those are measured. Further, as we have seen in Section 2 those may also be affected by seasonality. Lastly, there is a quadratic relationship to be modeled between temperature and salinity. These considerations lead to the following models.

In general we suppose that:

$$\mu_j(\mathbf{s},t) = \sum_{i=1}^{n_j} \beta_i^{(j)}\, u_i^{(j)}(\mathbf{s},t),\ j=1,2$$

where $u_i^{(j)}(\mathbf{s},t)$ is the value of the ith regressor observed at location $\mathbf{s}$ and at time t; n_j is the total number of regressors for the response $Z_j(\mathbf{s},t)$. Specifically, we assume that the mean level for temperature is given by:

$$\mu_1(\mathbf{s},t) = \beta_1^{(1)} + \beta_2^{(1)} s_1 + \beta_3^{(1)} s_2 + \beta_4^{(1)} s_1\, s_2 + \sum_{i=5}^{15} \beta_i^{(1)}\, u_{i-3}^{(1)}(\mathbf{s},t)$$

where the monthly seasonal indicators are given by

$$u_i^{(j)}(\mathbf{s},t) = \begin{cases} 1 & \text{if time } t \text{ is in the } i\text{th month} \\ 0 & \text{otherwise.} \end{cases}$$

The mean process for the salinity levels are modeled conditionally on temperature as follows:

$$\mu_2(\mathbf{s},t) = \beta_1^{(2)} + \beta_2^{(2)} s_1 + \beta_3^{(2)} s_2 + \beta_4^{(2)} s_1\, s_2 + \sum_{i=5}^{15} \beta_i^{(2)}\, u_{i-3}^{(2)}(\mathbf{s},t) + \beta_{16}^{(2)} z_1(\mathbf{s},t) + \beta_{17}^{(2)} z_1^2(\mathbf{s},t).$$

The above quadratic model has been justified previously in Section 2.

3.2 Kernel convolution effects

The spatio-temporal process $v_j(\mathbf{s},t)$, for both $j=1$ and 2 is thought to be induced by kernel convolution effects of a single latent spatio-temporal process $x(\boldsymbol{\omega},\tau)$ where $\boldsymbol{\omega}$ denotes a spatial location and τ denotes a time point. The same latent process $x(\boldsymbol{\omega},\tau)$ used in both $v_1(\mathbf{s},t)$ and $v_2(\mathbf{s},t)$ induces dependence between the data pairs $Z_1(\mathbf{s},t)$ and $Z_2(\mathbf{s},t)$.

Let $K_j(d_s,d_t), j=1,2$ denote the joint kernel in space and time where d_s and d_t are the distances in space and in time, respectively. Let $\boldsymbol{\omega}_l, l=1,\ldots,L$ denote the grid locations where the spatial smoothing kernels will be centered; similarly let $\tau_m, m=1,\ldots,M$ denote the equi-spaced time points where the temporal kernels will be centered. Now we write:

$$v_j(\mathbf{s},t) = \textstyle\sum_{l=1}^{L}\sum_{m=1}^{M} K_j(||\mathbf{s}-\boldsymbol{\omega}_l||, |t-\tau_m|)\, x(\boldsymbol{\omega}_l,\tau_m) \qquad (3)$$

where $||\mathbf{s} - \boldsymbol{\omega}_l||$ denotes the geodetic distance between the locations $\mathbf{s}$ and $\boldsymbol{\omega}_l$. In this paper we work with

$$K_j(d_s, d_t) = C_j(d_s)\, R_j(d_t)$$

where $C_j(d_s)$ is a kernel in space and $R_j(d_t)$ is a kernel in time for each $j = 1, 2$. Let $\phi_s^{(j)} > 0$, $\phi_t^{(j)} > 0$ denote the decay parameters in the jth spatial and temporal kernel respectively. Although other choices are possible we illustrate with

$$C_j(d_s) = \exp\{-\phi_s^{(j)} d_s\}, \text{ and } R_j(d_t) = \exp\{-\phi_t^{(j)} d_t\},$$

corresponding to exponential covariance functions. The parameters ϕ's determine the decay rate of the associated spatial and temporal correlations.

The latent process $x(\boldsymbol{\omega}, \tau)$ is assumed to have zero mean with a separable covariance structure (see e.g. Mardia and Goodall, 1993). That is,

$$\text{Cov}\big\{x(\boldsymbol{\omega}_l, \tau_m),\ x(\boldsymbol{\omega}_{l'}, \tau_{m'})\big\} = \sigma_x^2\ \rho_s(||\boldsymbol{\omega}_l - \boldsymbol{\omega}_{l'}||; \phi_{sx})\ \rho_t(||\tau_m - \tau_{m'}||; \phi_{tx}). \qquad (4)$$

In addition, the two ρ's are taken to be exponential covariance functions, i.e., $\rho(d; \phi) = \exp(-\phi d)$. After some preliminary investigation and tuning by using many runs of the Gibbs sampler we take $\phi_{tx} = 1$ and $\phi_{sx} = 0.001$ which correspond to an assumption of a smooth process with a spatial range of 3000 kilometers and a temporal range of 3 days approximately. (The range is defined as the approximate value of the distance, $d \approx -\log(0.05)/\phi$ where ϕ is the decay parameter.) These values provide adequate model validation, see Section 5.1 and imply a smooth latent process $x(\boldsymbol{\omega}, \tau)$.

Ideally, $\boldsymbol{\phi} = (\phi_s^{(1)}, \phi_t^{(1)}, \phi_s^{(2)}, \phi_t^{(2)})'$ should be estimated within the Bayesian model as well. However, in a classical inference setting it is not possible to consistently estimate all the parameters ϕ and σ^2 in a typical model for spatial data with a covariance function belonging to the Matèrn family, see Zhang (2004). Moreover, Stein (1999) shows that spatial interpolation is sensitive to the product $\sigma^2\phi$ but not to either one individually. In our Bayesian inference setup using Gibbs sampling joint estimation is often poorly behaved due to weak identifiability and extreme slow-mixing of the associated Markov chains under vague prior distributions for $\boldsymbol{\phi}$. In addition, the full conditional distribution for any of the decay parameters is not conjugate and sampling those in a Gibbs sampler requires expensive likelihood evaluations in each iteration. These difficulties are exacerbated by the large number of locations-time point combinations we work with here as well as the desire to do spatial prediction over the large rectangular box in Figure 1 covering most of the North Atlantic Ocean. In Section 5 we shall choose optimal values of $\boldsymbol{\phi}$ using a validation mean square error criterion and estimate the variances conditional on those values. Note that the full conditional distributions of the variances are conjugate under the assumption of conjugate prior distributions.

In our implementation we have taken $L = 36$ and $M = 12$, although we have experimented with other choices with both smaller and larger values. The points $\boldsymbol{\omega}_l$ are taken as the co-ordinates (latitude and longitude) of the grid points. The τ_m values are chosen to be 12 equi-distant time points between 1 to 365 days. These choices provided acceptable predictions and validations without making the MCMC algorithm too slow, see Section 5.1. The spatial locations of the 36 grid points are shown as triangles in Figure 1.

Let $\boldsymbol{\beta}$ denote the vector of unknown regression co-efficients $\beta_i^{(j)}, i = 1, \ldots, n_j, j = 1, 2$. Denote the unknown parameters by $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_x^2, \sigma_1^2, \sigma_2^2)'$. We assume that, a priori, the β's are independent with distribution $N(0, A^2)$. We take A^2 to be large for vague prior specification. For the three variance parameters σ_x^2, σ_1^2, and σ_2^2 we assume independent inverse gamma prior distributions, $IG(a, b)$ (with mean $b/(a-1)$) setting $a = 2$ and $b = 1$ to have proper prior distributions with mean 1 and infinite variance. Let $\pi(\boldsymbol{\theta})$ denote the product of the prior densities for $\boldsymbol{\beta}$, σ_x^2, σ_1^2, and σ_2^2. The joint posterior distribution is the product of the likelihood and prior specifications and is provided in the appendix.

4 Prediction details

4.1 Prediction at any given time point

The modeling in Section 3 allows us to interpolate the spatial surface at any time point t' which can be in the past or the future. More precisely, using (1) and (2), for a new location $\mathbf{s}'$ at time t', $Z_j(\mathbf{s}', t')$ is conditionally independent of $\mathbf{z}$ given $v_j(\mathbf{s}', t')$ with its distribution given by

$$Z_j(\mathbf{s}', t') \sim N\left(\mu_j(\mathbf{s}', t') + v_j(\mathbf{s}', t'),\ \sigma_j^2\right). \tag{5}$$

The posterior predictive distribution of $Z_j(\mathbf{s}', t')$ is obtained by integrating over the unknown parameters in (5) with respect to the joint posterior distribution, that is,

$$\pi\left(Z_j(\mathbf{s}', t')|\mathbf{z}\right) = \int \pi\left(Z_j(\mathbf{s}', t')|\mathbf{x}, \boldsymbol{\theta}\right) \pi\left(\boldsymbol{\theta}, \mathbf{x}|\mathbf{z}\right)\, d\mathbf{x}\, d\boldsymbol{\theta}, j = 1, 2, \tag{6}$$

where $\mathbf{x}$ denote the collection of all the $x(\boldsymbol{\omega}_l, \tau_m)$ values. When using MCMC methods to draw samples from the posterior, the predictive distribution (6) is sampled by composition; draws from the posterior, $\pi(\boldsymbol{\theta}, \mathbf{x}|\mathbf{z})$ enable draws for $Z_1(\mathbf{s}', t')$ and subsequently $Z_2(\mathbf{s}', t')$.

4.2 Annual predictions

It is of interest to develop methodology for average annual prediction surfaces both for temperature and salinity. The annual predictions are to be obtained using the predictive distributions similar to(6) with the following modifications. The annual averages should be obtained by averaging the model for daily data, specifically for a new location $\mathbf{s}'$ the annual average is given by:

$$Z_j(\mathbf{s}') = \frac{1}{T} \sum_{t=1}^{T} Z_j(\mathbf{s}', t)$$

where T is the number of days in a year, either 365 or 366 depending on whether the year is a leap year. From model (5) we have that:

$$Z_j(\mathbf{s}') \sim N\left(\bar{\mu}_j(\mathbf{s}') + \bar{v}_j(\mathbf{s}'),\ \frac{\sigma_j^2}{T}\right) \tag{7}$$

where

$$\bar{\mu}_j(\mathbf{s}') = \frac{1}{T}\sum_{t=1}^{T}\mu_j(\mathbf{s}',t) \text{ and } \bar{v}_j(\mathbf{s}') = \frac{1}{T}\sum_{t=1}^{T} v_j(\mathbf{s}',t).$$

Note that

$$\begin{aligned}\bar{v}_j(\mathbf{s}') &= \tfrac{1}{T}\textstyle\sum_{t=1}^{T} v_j(\mathbf{s}',t)\\ &= \tfrac{1}{T}\textstyle\sum_{t=1}^{T}\sum_{l=1}^{L}\sum_{m=1}^{M} K_j(||\mathbf{s}-\boldsymbol{\omega}_l||, |t-\tau_m|)\, x(\boldsymbol{\omega}_l,\tau_m)\\ &= \textstyle\sum_{l=1}^{L}\sum_{m=1}^{M} C_j(||\mathbf{s}-\boldsymbol{\omega}_l||)\, x(\boldsymbol{\omega}_l,\tau_m)\tfrac{1}{T}\sum_{t=1}^{T} R_j(|t-\tau_m|).\end{aligned}$$

Thus it is straightforward to calculate $\bar{v}_j(\mathbf{s}')$ for $j = 1, 2$. The calculation of $\bar{\mu}_1(\mathbf{s}')$ is also straightforward, since it is given by:

$$\bar{\mu}_j(\mathbf{s}') = \beta_1^{(1)} + \beta_2^{(1)} s_1 + \beta_3^{(1)} s_2 + \beta_4^{(1)} s_1\, s_2 + \sum_{i=5}^{15}\beta_i^{(1)}\,\bar{u}_i^{(1)}(\mathbf{s}'),$$

where $\bar{u}_i^{(1)}(\mathbf{s}') = \frac{1}{T}\sum_{t=1}^{T} u_i^{(1)}(\mathbf{s}',t)$. Now

$$\bar{\mu}_2(\mathbf{s}') = \beta_1^{(2)} + \beta_2^{(2)} s_1 + \beta_3^{(2)} s_2 + \beta_4^{(2)} s_1\, s_2 + \sum_{i=5}^{15}\beta_i^{(2)}\,\bar{u}_{i-3}^{(2)}(\mathbf{s}) + \beta_{16}^{(2)}\bar{z}_1(\mathbf{s}') + \beta_{17}^{(2)}\frac{1}{T}\sum_{t=1}^{T} z_1^2(\mathbf{s}',t).$$

The quantity $\sum_{t=1}^{T} z_1^2(\mathbf{s}',t)$ can only be calculated exactly, if we have the predictions for each $\mathbf{s}'$ and $t = 1,\ldots,T$. However, due to the large number of spatial prediction locations $\mathbf{s}'$ and large value of T this is a huge computational burden. Instead, following Sahu *et al.* (2006) we shall use an approximation. In particular, we shall use

$$\frac{1}{T}\sum_{t=1}^{T}\left\{z_1(\mathbf{s}',t) - \bar{z_1}(\mathbf{s}')\right\}^2 \approx \frac{1}{B}\sum_{j=1}\left\{z_1(\mathbf{s}',t_j) - \bar{z}_{1B}(\mathbf{s}')\right\}^2$$

where B is a number much less than T and the sequence t_j is an equally spaced subsequence of $\{1,\ldots,T\}$ and $\bar{z}_{1B}(\mathbf{s}') = \sum_{j=1}^{B} z_1(\mathbf{s}',t_j)/B$. In our implementation we take $B = 4$ and take the four days equally spaced in the year. See Sahu *et al.* (2006) for more details in this regard.

Once the parameters in (7) have been sampled we use the predictive distribution, analogous to (6), given by:

$$\pi\left(Z_j(\mathbf{s}')|\mathbf{z}\right) = \int \pi\left(Z_j(\mathbf{s}')|\mathbf{x},\boldsymbol{\theta}\right)\pi\left(\boldsymbol{\theta},\mathbf{x}|\mathbf{z}\right)\, d\mathbf{x}\, d\boldsymbol{\theta}, j = 1,2.$$

5 Summary of analysis

5.1 Model choice and validation

As mentioned in Section 3 the four decay parameters in $\boldsymbol{\phi} = (\phi_s^{(1)},\ \phi_t^{(1)}, \phi_s^{(2)},\ \phi_t^{(2)})'$ are to be chosen by cross validation. For this we consider the validation mean-square error

$$\text{VMSE} = \frac{1}{250}\sum_{p=1}^{250}\left(Z_p - \hat{Z}_p\right)^2$$

where Z_p and $\hat{Z}_p$ denote the pth datum set aside for model validation purposes (see Section 2) and its mean predicted value respectively.

A full search for the four dimensional optimal value of $\boldsymbol{\phi}$ within a grid of any reasonable size for three different models (surface, mid-layer and deep) is computationally prohibitive. From many preliminary Gibbs sampling run of the models, we have found that the VMSE is not sensitive to changes in the temporal decay parameters, $\phi_t^{(1)}$ and $\phi_t^{(2)}$ when those are fixed near 1. Henceforth, we adopt this value which correspond to a temporal range of approximately three days since $\exp(-3) \approx 0.05$.

For the remaining two parameters, $\phi_s^{(1)}$ and $\phi_s^{(2)}$, we search for optimal values so that the ranges are in between 300 to 6,500 kilometers. The optimal values and the associated ranges in kilometers are presented in Table 1. For the surface data, the optimal value of the range for temperature is about 375 kilometers while for salinity it is about 600 kilometers. This may seem to be surprising since there exist non-linear relationships between ocean temperature and salinity. However, the relationships are not one-to-one and for observed data the exact relationships may not hold as the observations may be noisy and are not collected always at 'laboratory conditions'. The optimal range values for the other two layers become large as depth increases and are reported in the last two rows of Table 1. These values are intuitively sensible as the ocean characteristics changes very slowly in the mid and deep ocean. Henceforth we work with these optimal values.

We now turn to validate the models for three different depths with the optimal spatial decay parameters chosen above. Recall that we have set aside 250 observations (temperature and salinity values) from each of three sets of data. Figure 4 plots the validation 95% prediction intervals as vertical bars and actual observations as points. A broken vertical line implies that the observation is not contained within the prediction interval. The proportions of 250 validation observations contained within the prediction intervals are labeled on the plot as well. These are also given in Table 2. The proportions are in the range 0.92 and 0.97 and show a better coverage provided by salinity intervals. This is expected since salinity varies less than temperature. Overall, this shows that the models are performing adequately for out of sample validated predictions.

5.2 Parameter estimates

We first consider the results for the surface data. Here the seasonal components are included both for temperature and salinity. The parameter estimates for temperature are given in Table 3 and those for salinity are given in Table 4. As expected, ocean temperatures are seen to be significantly cooler with less salinity at higher latitudes. The same conclusion holds for higher longitudes as well but this effect is somewhat negated by the positively significant interaction parameter. There is more variability in the temperature than the salinity levels as seen by the estimates of the σ^2's.

The parameter estimates of monthly seasonal components agree with the patterns seen in Figure 3, namely temperatures in July–October are higher than the remaining months. The salinity levels do not vary a great deal from month to month, although the months of August, September and October are the months with lowest salinity levels

after adjusting for temperature.

The parameter estimates for the mid-layer data set are given in Tables 5 and 6 and those for the deep layer data set are given in Tables 7 and 8. As mentioned before, the monthly seasonal terms have not been included for these two layers. The effect of latitude and longitude on salinity levels gets considerably weaker at higher depths. However, these effects continue to remain significant for temperature at higher depths. The effect of longitude at deep layer is seen to be positive where the interaction parameter between latitude and longitude is not present since this was not significant. There is less variability in temperature at the deep layer than the other two layers, as expected.

5.3 Prediction

As mentioned before we have only a few data points to obtain prediction maps for a particular day. The daily prediction maps will have a large amount of uncertainty associated with them. Besides, we have validated a large number of site-wise daily predictions in all three layers of the ocean. That is why we do not report the maps of daily prediction, instead we turn to the annual summaries. The annual predictions of mean temperature and salinity are reported in Figures 5, 6 and 7. The predictions show two distinct ocean currents: the cooler polar currents and the warmer equatorial currents. The predictions for the deep ocean, however, show that the two ocean characteristics are much less variable at this depth, as would be expected. The associated uncertainty maps are also provided in the figures. In general these map show that there is less variability in predictions at the sites near to the sampling locations. Also as expected, variability increases for the prediction sites which are further away from the sampling sites, see the standard deviation maps around the three corners, south-west, south-east and north-east where there are few or no sampling sites, see Figure 1 as well.

We have not found comparable annual prediction maps from any other project. The Coriolis project (http://www.coriolis.eu.org/) produces a ten-day near-real time analysis of the data. Their system uses an objective analysis scheme. In essence this involves kriging on residuals from a prior mean. This is taken from the world ocean atlas 2001. In contrast to the results presented here they do not take into account the temporal aspects of the problem producing a separate analysis for each ten-day period.

6 Discussion

In this paper we have formulated joint models for temperature and salinity levels observed at three different depths of the North Atlantic. We have shown how to use this model to obtain annual temperature and salinity maps along with the associated uncertainty maps. The empirical model based techniques have been adequately verified by validating a large number of held out data. The empirical models are useful since salinity cannot yet be measured by satellites, and space-based measurements can only ever give us values at the surface. In future work, we plan to investigate a joint model capturing the space-time variation in the joint relationships between temperature and salinity levels at all three depths. Recently developed anisotropic and non-stationary

models see, e.g. Fuentes and Smith (2001), Schmidt and O'Hagan (2003), Pintore and Holmes (2004).

Acknowledgements

The Argo data used in this paper were collected and made freely available by the International Argo Project and the national programs that contribute to it (www.argo.ucsd.edu and argo.jcommops.org). Argo is a pilot program of the Global Ocean Observing System.

Appendix: Joint and conditional posterior distributions

Let us define

$$X_{M\times L} = \begin{pmatrix} x_{\boldsymbol{\omega}_1,\tau_1} & x_{\boldsymbol{\omega}_2,\tau_1} & \cdots & x_{\boldsymbol{\omega}_L,\tau_1} \\ x_{\boldsymbol{\omega}_1,\tau_2} & x_{\boldsymbol{\omega}_2,\tau_2} & \cdots & x_{\boldsymbol{\omega}_L,\tau_2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{\boldsymbol{\omega}_1,\tau_M} & x_{\boldsymbol{\omega}_2,\tau_M} & \cdots & x_{\boldsymbol{\omega}_L,\tau_M} \end{pmatrix}.$$

We concatenate the columns of the matrix X to obtain the vector $\mathbf{x}$. Let Σ_{sx} and Σ_{tx} denote the spatial and temporal correlation matrices of the $x(\boldsymbol{\omega},\tau)$ process. That is, for $l, l' = 1,\ldots,L$ and $m, m' = 1,\ldots,M$, we have:

$$(\Sigma_{sx})_{ll'} = \rho_{sx}(\boldsymbol{\omega}_l - \boldsymbol{\omega}_{l'};\phi_{sx}), \quad (\Sigma_{tx})_{mm'} = \rho_{tx}(m-m';\phi_{tx}).$$

Now the prior specification for the $x(\boldsymbol{\omega},\tau)$ process, $\pi(\mathbf{x}|\sigma_x^2)$, is seen to be

$$\mathbf{x} \sim N\left(\mathbf{0}, \sigma_x^2\,\Sigma_{sx}\otimes\Sigma_{tx}\right)$$

where $\otimes$ denotes the Kronecker product.

The log-likelihood is written as:

$$l(\boldsymbol{\theta},\mathbf{x};\mathbf{z}) = -\frac{N}{2}\log(\sigma_1^2) - \frac{N}{2}\log(\sigma_2^2) - \frac{1}{2}\sum_{j=1}^{2}\frac{1}{\sigma_j^2}\sum_{p=1}^{N}\left\{z_j(\mathbf{s},t) - \mu_j(\mathbf{s},t) - v_j(\mathbf{s},t)\right\}^2 + C$$

where $\mathbf{z}$ denotes all the data and C is a constant free of $\boldsymbol{\theta}$ and $\mathbf{x}$. In the above log-likelihood recall that we use p to index a space-time combination denoted by $\mathbf{s}$ and t.

The joint posterior distribution is now given by:

$$\pi(\boldsymbol{\theta},\mathbf{x}|\mathbf{z}) \propto \exp\left[l(\boldsymbol{\theta},\mathbf{x};\mathbf{z})\right]\,\pi(\mathbf{x}|\sigma_x^2)\,\pi(\boldsymbol{\theta}).$$

The complete conditional distributions needed for Gibbs sampling are derived from $\pi(\boldsymbol{\theta},\mathbf{x}|\mathbf{z})$ and are given below. Let us define

$$\mathbf{K}_{jp}^{(s)'} = \left(C_j(d(\mathbf{s}_p,\boldsymbol{\omega}_1)),\ldots,C_j(d(\mathbf{s}_p,\boldsymbol{\omega}_L))\right),$$

$$\mathbf{K}_{jp}^{(t)'} = \left(R_j(d(t_p,\tau_1)),\ldots,R_j(d(t_p,\tau_M))\right),$$

for $j = 1, 2$ and $p = 1,\ldots,N$. For each $j = 1, 2$ define the $N \times ML$ matrix:

$$K_j = \begin{pmatrix} \left(\mathbf{K}_{j1}^{(s)} \otimes \mathbf{K}_{j1}^{(t)}\right)' \\ \vdots \\ \left(\mathbf{K}_{jp}^{(s)} \otimes \mathbf{K}_{jp}^{(t)}\right)' \\ \vdots \\ \left(\mathbf{K}_{jN}^{(s)} \otimes \mathbf{K}_{jN}^{(t)}\right)' \end{pmatrix}.$$

Now Equation (3) is equivalently written as:

$$\mathbf{v}_j = K_j\mathbf{x}. \tag{8}$$

The log-likelihood contribution for $\mathbf{x}$ is now given by:

$$-\frac{1}{2}\sum_{j=1}^{2}\frac{1}{\sigma_j^2}\left(\mathbf{z}_j - \boldsymbol{\mu}_j - K_j\mathbf{x}\right)'\left(\mathbf{z}_j - \boldsymbol{\mu}_j - K_j\mathbf{x}\right)$$

where $\mathbf{z}_j$ and $\boldsymbol{\mu}_j$ are vectors of order N with elements $z_j(\mathbf{s},t)$ and $\mu_j(\mathbf{s},t)$. Therefore, the complete conditional distribution for sampling $\mathbf{x}$ is normal with mean $\boldsymbol{\xi}$ and covariance Λ where

$$\boldsymbol{\xi} = \Lambda\sum_{j=1}^{2}\frac{1}{\sigma_j^2}K_j'(\mathbf{z}_j - \boldsymbol{\mu}_j) \text{ and covariance } \Lambda^{-1} = \frac{1}{\sigma_x^2}\Sigma_{sx}^{-1}\otimes\Sigma_{tx}^{-1} + \sum_{j=1}^{2}\frac{1}{\sigma_j^2}K_j'K_j.$$

The complete conditional distribution of $\beta_i^{(j)}$ is normal with mean

$$\zeta_i^{(j)}\left[\frac{1}{\sigma_j^2}\sum_{p=1}^{N}u_i^{(j)}(\mathbf{s},t)\left\{z_j(\mathbf{s},t) - \mu_j(\mathbf{s},t) + \beta_i^{(j)}u_i^{(j)}(\mathbf{s},t) - v_j(\mathbf{s},t)\right\}\right],$$

and variance

$$\zeta_i^{(j)} = \left(\frac{1}{\sigma_j^2}\sum_{p=1}^{N}u_i^{(j)}(\mathbf{s},t)^2 + \frac{1}{A^2}\right)^{-1}.$$

This sampling scheme updates the components of $\boldsymbol{\beta}$ one after another. However, $\boldsymbol{\beta}$ can be sampled in a single block as well to obtain faster convergence at additional programming cost.

We also obtain the following complete conditional distributions by straightforward calculations:

$$\begin{aligned} &\tfrac{1}{\sigma_j^2} \sim G\left(\tfrac{N}{2} + a, b + \tfrac{1}{2}\textstyle\sum_{p=1}^{N}\left\{z_j(\mathbf{s},t) - \mu_j(\mathbf{s},t) - v_j(\mathbf{s},t)\right\}^2\right), \\ &\tfrac{1}{\sigma_x^2} \sim G\left(\tfrac{LM}{2} + a, b + \mathbf{x}'\Sigma_{sx}^{-1}\otimes\Sigma_{tx}^{-1}\mathbf{x}\right), \end{aligned}$$

where $G(a,b)$ denotes the gamma distribution with mean a/b.

REFERENCES

1. Ferrari, R. and Polzin, K. L. 2005. Finescale structure of the T-S relation in the eastern North Atlantic. *Journal of Physical Oceanography*, **35**, 1437–1454.

2. Ferrari, R. and Rudnick, D. L. 2000. Thermohaline variability in the upper ocean. *Journal of Geophysical Research-Oceans*, **105** (C7): 16857-16883.

3. Fuentes, M. and Smith, R. L. 2001. A new class of non-stationary spatial models. Technical report, Department of Statistics, North Carolina state University.

4. Gelfand A. E., Banerjee, S. and Gamerman, D. 2005. Spatial process modelling for univariate and multivariate dynamic spatial data *Environmetrics*, **16**, 465–479.

5. Gelfand, A. E., Schmidt, A. M., Banerjee, S. and Sirmans, C. F. 2004. Nonstationary Multivariate Process Modelling through Spatially Varying Coregionalization (with discussion). *Test*, **2**, 1–50.

6. Higdon, D. M. 1998. A process-convolution approach to modeling temperatures in the north Atlantic Ocean. *Journal of Envioronmental and Ecological Statistics*, **5**, 173–190.

7. Mardia, K. V. and Goodall, C., 1993. Spatial-temporal analysis of multivariate environmental monitoring data. In *Multivariate Environmental Statistics*. (Eds. G. P. Patil and C. R. Rao). Amsterdam: Elsevier, pp 347–386.

8. Pintore, A. and Holmes, C. 2004. Non-stationary covariance functions via spatially adaptive spectra. to appear in *J. Amer. Statist. Assoc.*

9. Sahu, S. K. and Mardia, K. V. 2005a. Recent Trends in Modeling Spatio-Temporal Data. In Proceedings of the special meeting on Statistics and Environment organized by the Società Italiana di Statistica held in Università Di Messina, September 21-23, 2005, Invited Papers, pages 69–83. Published by the Università Di Messina, Messina, Italy.

10. Sahu, S. K. and Mardia, K. V. 2005b. A Bayesian Kriged-Kalman model for short-term forecasting of air pollution levels. *Journal of the Royal Statistical Society*, Series C, **54**, 223–244.

11. Sahu, S.K., Gelfand, A. E. and Holland, D. M. 2006. Spatio-temporal modeling of fine particulate matter. *Journal of Agricultural, Biological, and Environmental Statistics,* **11**, 61–86.

12. Sahu, S. K., Gelfand, A. E. and Holland, D. M. 2007. High Resolution Space-Time Ozone Modeling for Assessing Trends. To appear in *Journal of the American Statistical Association.*

13. Schmidt, A. and O'Hagan, A. 2003. Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *J. Roy. Stat. Soc.*, B, **65**, 743–758.

14. Stein, M. 1999. *Interpolation of Spatial Data: Some Theory for Kriging:* Springer Verlag.

15. Ver Hoef, J. M. and Barry, R. D. 1998. Constructing and fitting models for cokriging and multivariate spatial prediction. *Journal of Statistical Planning and Inference*, **69**, 275–294.

16. Wikle, C. K. 2003. Hierarchical models in environmental science. *International Statistical Review*, **71**, 181–199.

17. Zhang, H. 2004. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, **99**, 250–261.

Table 1: The optimal decay parameters and the associated approximate range values in kilometers.

	Temperature		Salinity	
	$\phi_s^{(1)}$	Range	$\phi_s^{(2)}$	Range
Surface	0.008	375	0.005	600
Mid-Layer	0.004	750	0.003	1000
Deep	0.0005	6000	0.001	3000

Table 2: The proportion of 250 validation observations lying within the 95% prediction intervals.

	Temperature	Salinity
Surface	0.94	0.97
Mid-Layer	0.92	0.94
Deep	0.95	0.97

Table 3: The estimates of the parameters for surface temperature.

	mean	sd	95% interval
$\beta_1^{(1)}$	23.043	0.781	(21.547, 24.538)
$\beta_2^{(1)}$ (latitude)	–0.152	0.012	(–0.179, –0.134)
$\beta_3^{(1)}$ (longitude)	–0.439	0.016	(–0.463, –0.407)
$\beta_4^{(1)}$ (interaction)	0.011	0.0004	(0.010, 0.011)
$\beta_5^{(1)}$ (Feb)	–1.353	0.597	(–2.618, –0.255)
$\beta_6^{(1)}$ (Mar)	–1.777	0.506	(–2.909, –0.850)
$\beta_7^{(1)}$ (Apr)	–1.636	0.507	(–2.768, –0.712)
$\beta_8^{(1)}$ (May)	–0.841	0.506	(–1.985, 0.085)
$\beta_9^{(1)}$ (Jun)	0.870	0.508	(–0.279, 1.795)
$\beta_{10}^{(1)}$ (Jul)	2.744	0.506	(1.582, 3.666)
$\beta_{11}^{(1)}$ (Aug)	4.137	0.503	(2.996, 5.053)
$\beta_{12}^{(1)}$ (Sep)	4.086	0.504	(2.953, 4.992)
$\beta_{13}^{(1)}$ (Oct)	2.870	0.504	(1.728, 3.787)
$\beta_{14}^{(1)}$ (Nov)	1.072	0.506	(–0.088, 1.998)
$\beta_{15}^{(1)}$ (Dec)	–0.053	0.649	(–1.189, 0.873)
σ_ϵ	1.699	0.026	(1.649, 1.752)
σ_x	0.847	0.372	(0.404, 1.907)

Table 4: The estimates of the parameters for surface salinity.

	mean	sd	95% interval
$\beta_1^{(2)}$	35.982	0.151	(35.707, 36.316)
$\beta_2^{(2)}$ (latitude)	–0.004	0.003	(–0.010, 0.002)
$\beta_3^{(2)}$ (longitude)	–0.001	0.006	(–0.011, 0.013)
$\beta_4^{(2)}$ (interaction)	0.001	0.0001	(0.0, 0.0006)
$\beta_5^{(2)}$ (Feb)	–0.021	0.122	(–0.287, 0.198)
$\beta_6^{(2)}$ (Mar)	–0.082	0.102	(–0.311, 0.076)
$\beta_7^{(2)}$ (Apr)	–0.083	0.102	(–0.315, 0.073)
$\beta_8^{(2)}$ (May)	–0.173	0.102	(–0.406, –0.020)
$\beta_9^{(2)}$ (Jun)	–0.358	0.103	(–0.600, –0.202)
$\beta_{10}^{(2)}$ (Jul)	–0.592	0.104	(–0.838, –0.435)
$\beta_{11}^{(2)}$ (Aug)	–0.789	0.104	(–1.039, –0.636)
$\beta_{12}^{(2)}$ (Sep)	–0.837	0.104	(–1.088, –0.685)
$\beta_{13}^{(2)}$ (Oct)	–0.701	0.103	(–0.943, –0.548)
$\beta_{14}^{(2)}$ (Nov)	–0.448	0.102	(–0.689, –0.292)
$\beta_{15}^{(2)}$ (Dec)	–0.313	0.102	(–0.543, –0.161)
$\beta_{16}^{(2)}$ (Temp)	–0.011	0.008	(–0.024, 0.004)
$\beta_{17}^{(2)}$ (Temp^2)	0.004	0.0002	(0.0034, 0.0044)
σ_ϵ	0.355	0.006	(0.344, 0.366)

Table 5: The estimates of the parameters for mid-layer temperature.

	mean	sd	95% interval
$\beta_1^{(1)}$	23.357	1.120	(21.873, 25.997)
$\beta_2^{(1)}$ (latitude)	–0.160	0.019	(–0.209, –0.133)
$\beta_3^{(1)}$ (longitude)	–0.429	0.025	(–0.465, –0.364)
$\beta_4^{(1)}$ (interaction)	0.011	0.001	(0.009, 0.012)
σ_ϵ	1.795	0.025	(1.746, 1.845)
σ_x	0.938	0.312	(0.459, 1.711)

Table 6: The estimates of the parameters for mid-layer salinity

	mean	sd	95% interval
$\beta_1^{(2)}$	35.128	0.039	(35.037, 35.185)
$\beta_2^{(2)}$ (latitude)	–0.0011	0.0006	(–0.002, 0.0001)
$\beta_3^{(2)}$ (longitude)	0.004	0.0007	(0.0026, 0.0051)
$\beta_4^{(2)}$ (Temp)	–0.040	0.0074	(–0.047, –0.019)
$\beta_5^{(2)}$ (Temp^2)	0.0076	0.0004	(0.007, 0.008)
σ_ϵ	0.052	0.0008	(0.051, 0.054)

Table 7: The estimates of the parameters for deep-layer temperature.

	mean	sd	95% interval
$\beta_1^{(1)}$	15.228	0.117	(15.000, 15.471)
$\beta_2^{(1)}$ (latitude)	–0.120	0.002	(–0.124, -0.116)
$\beta_3^{(1)}$ (longitude)	0.119	0.002	(0.115, 0.124)
σ_ϵ	1.147	0.016	(1.115, 1.180)
σ_x	0.665	0.188	(0.404, 1.135)

Table 8: The estimates of the parameters for deep-layer salinity.

	mean	sd	95% interval
$\beta_1^{(2)}$	34.931	0.038	(34.823, 34.990)
$\beta_2^{(2)}$ (latitude)	0.002	0.0003	(0.0018, 0.003)
$\beta_3^{(2)}$ (longitude)	0.003	0.0002	(0.0026, 0.0036)
$\beta_4^{(2)}$ (Temp)	–0.065	0.006	(–0.074, –0.048)
$\beta_5^{(2)}$ (Temp^2)	0.015	0.0004	(0.014, 0.016)
σ_ϵ	0.067	0.001	(0.065, 0.069)

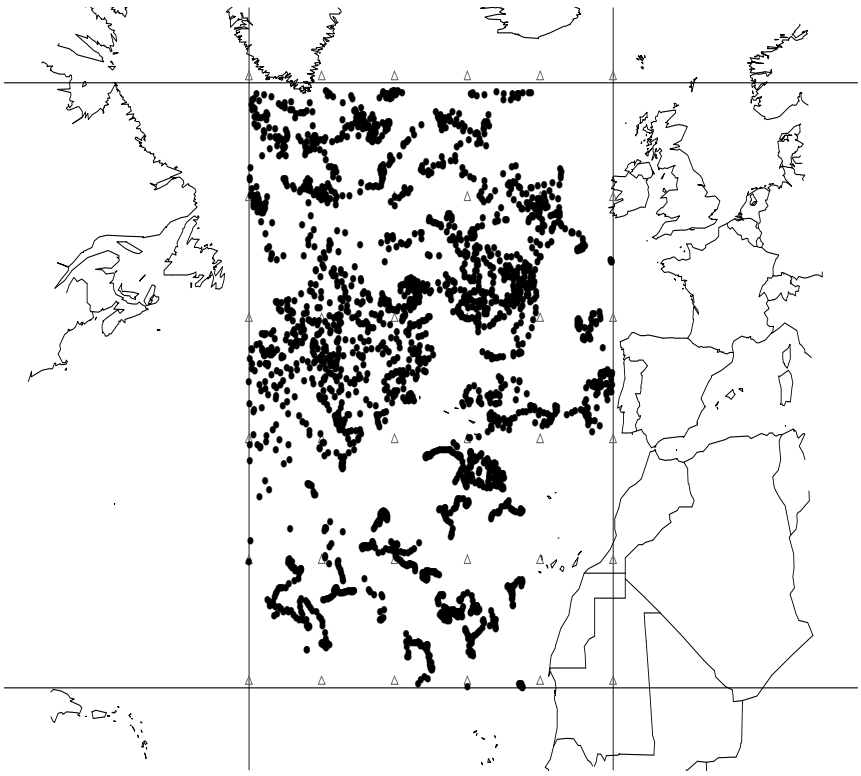

Figure 1: A map of the North Atlantic Ocean showing the 2374 observation locations at the surface layer. The spatial locations of 36 grid points are plotted as triangles.

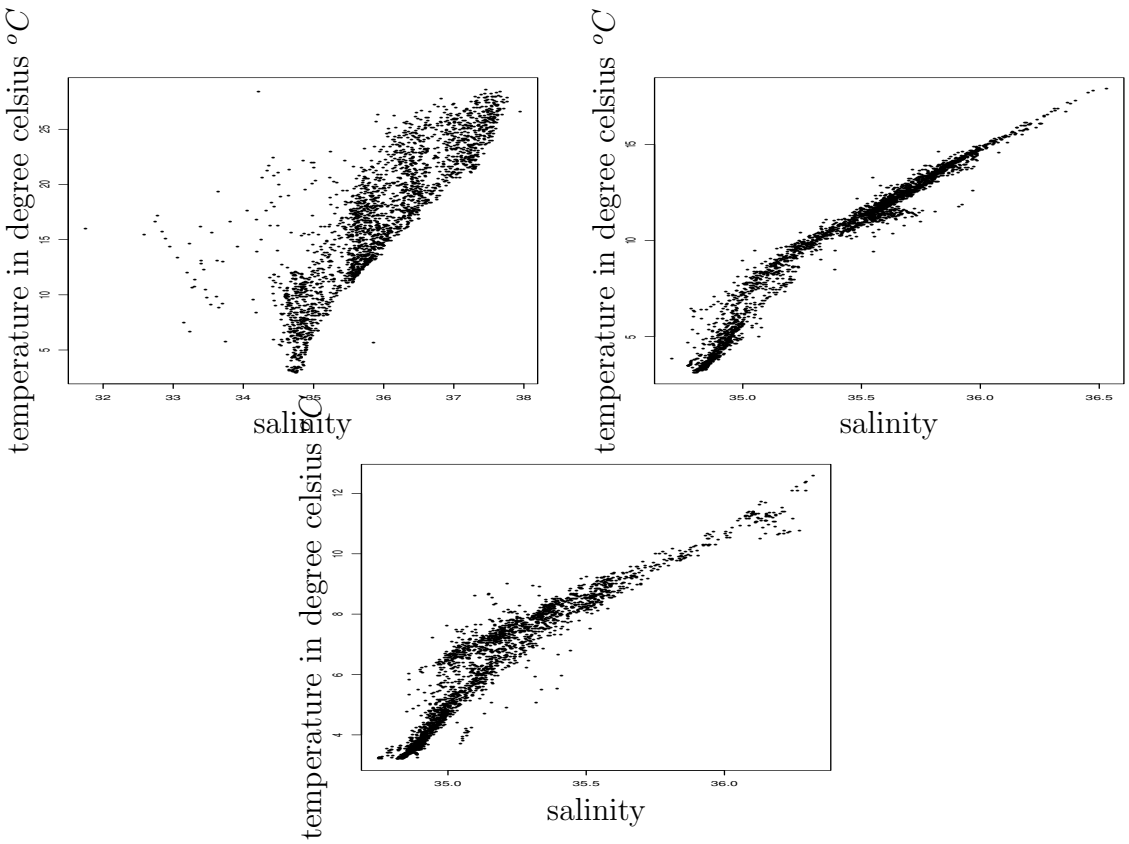

Figure 2: Scatter plots of temperature and salinity at: the surface (top panel); mid-layer (middle panel) and the deep ocean (bottom panel).

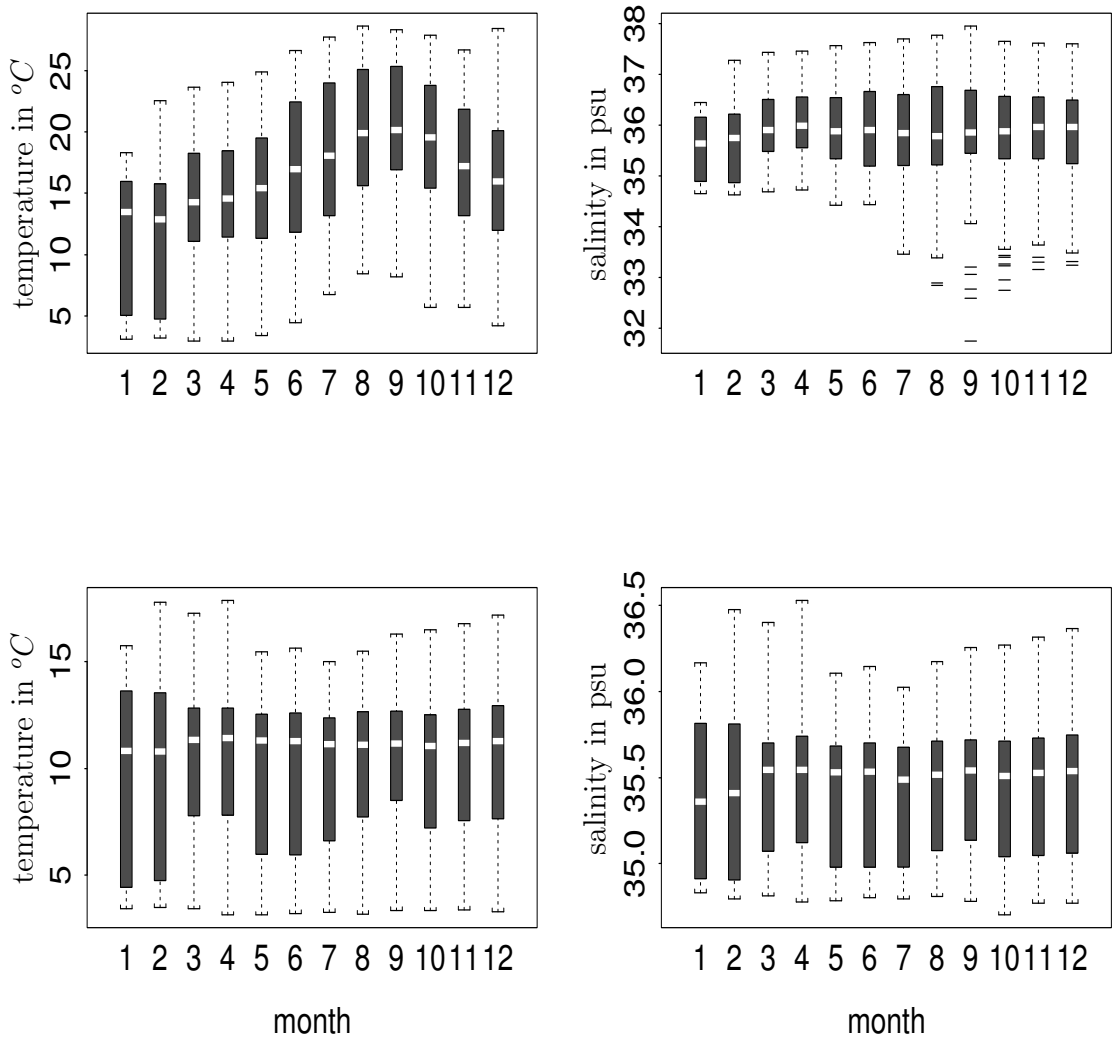

Figure 3: The boxplot of the temperature and salinity levels for 12 months, at the surface (first row) and at the mid-layer (second row). psu stands for practical salinity units.

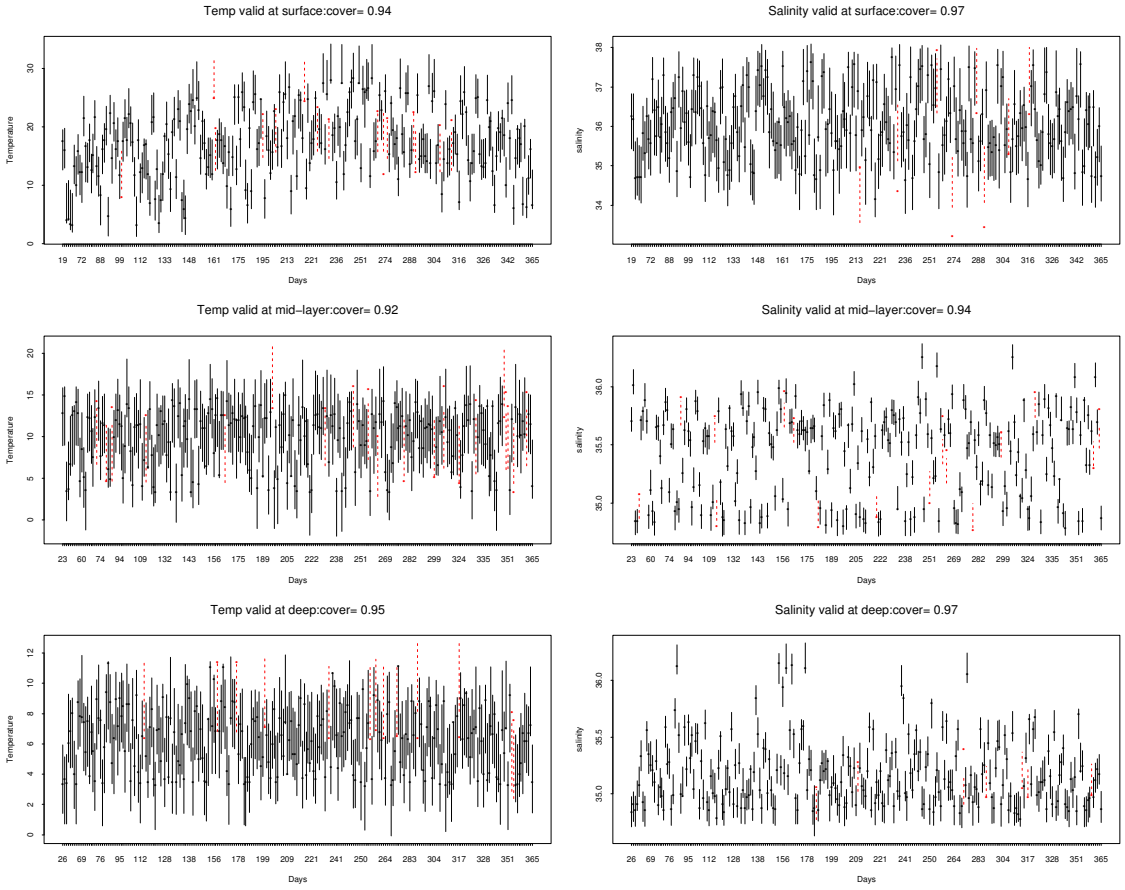

Figure 4: Validation plots at the three layers. The vertical bars represent 95% prediction intervals and the actual observations are shown as points. A broken vertical bar shows that the corresponding observation is outside the prediction interval. The proportions of 250 validation observations contained within the prediction intervals are also labelled on the plot.

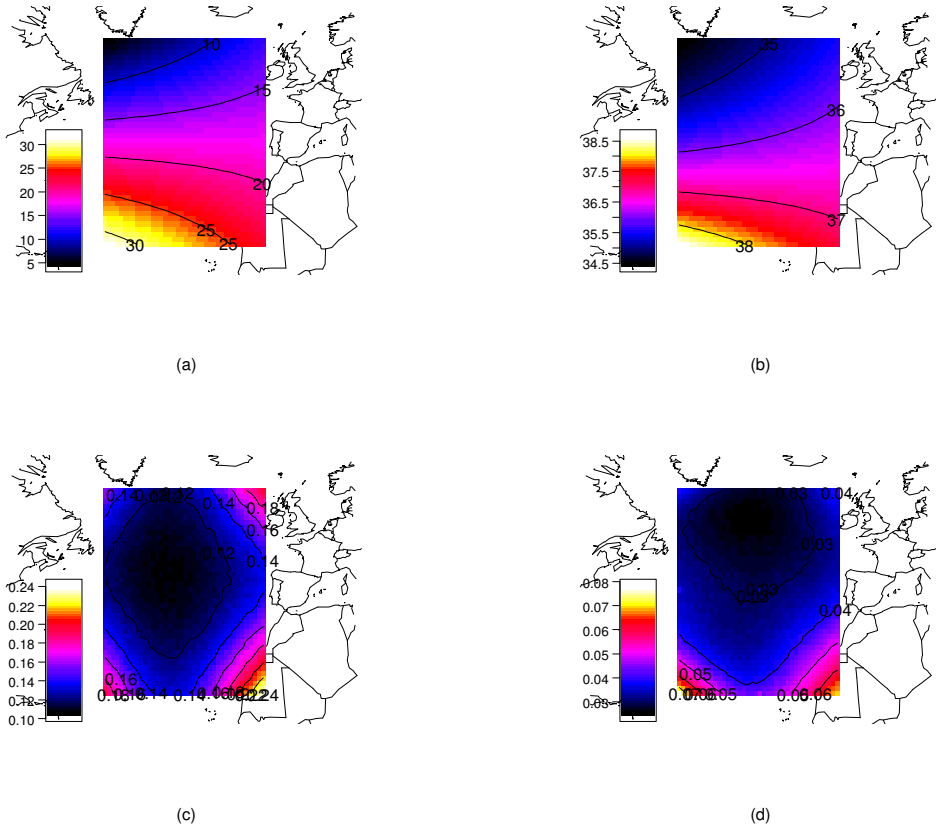

Figure 5: Annual prediction maps and their standard deviations at the surface: temperature on panel (a); salinity on panel (b); standard deviation of temperature on panel (c); standard deviation of salinity on panel (d).

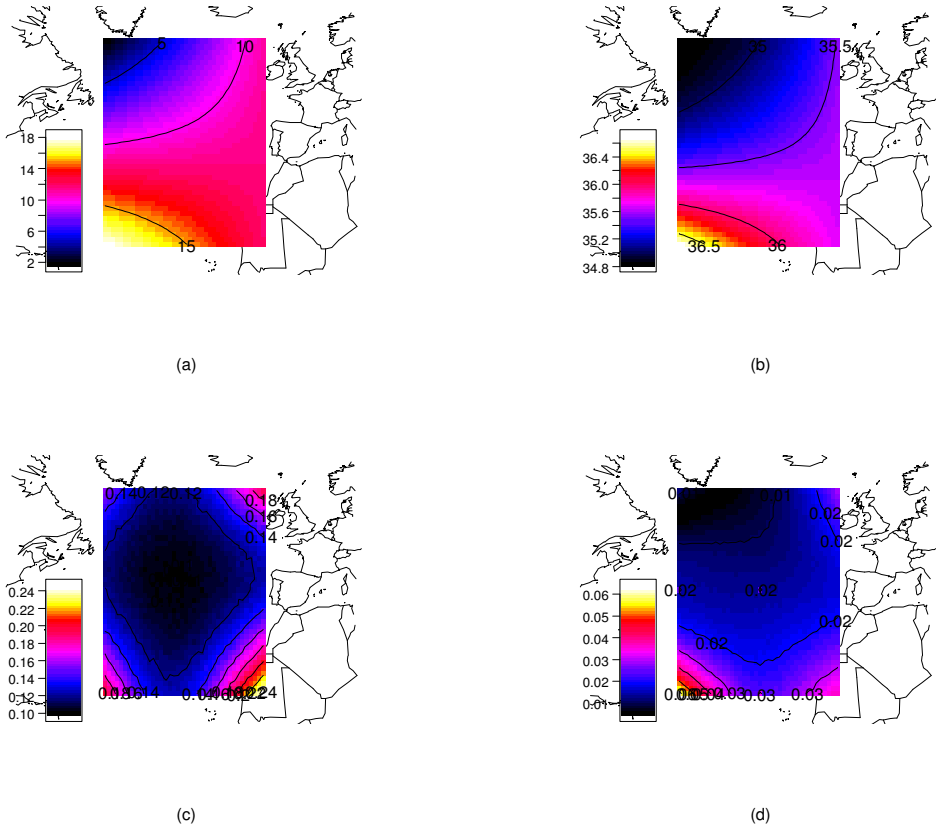

Figure 6: Annual prediction maps and their standard deviations at the mid-layer: temperature on panel (a); salinity on panel (b); standard deviation of temperature on panel (c); standard deviation of salinity on panel (d).

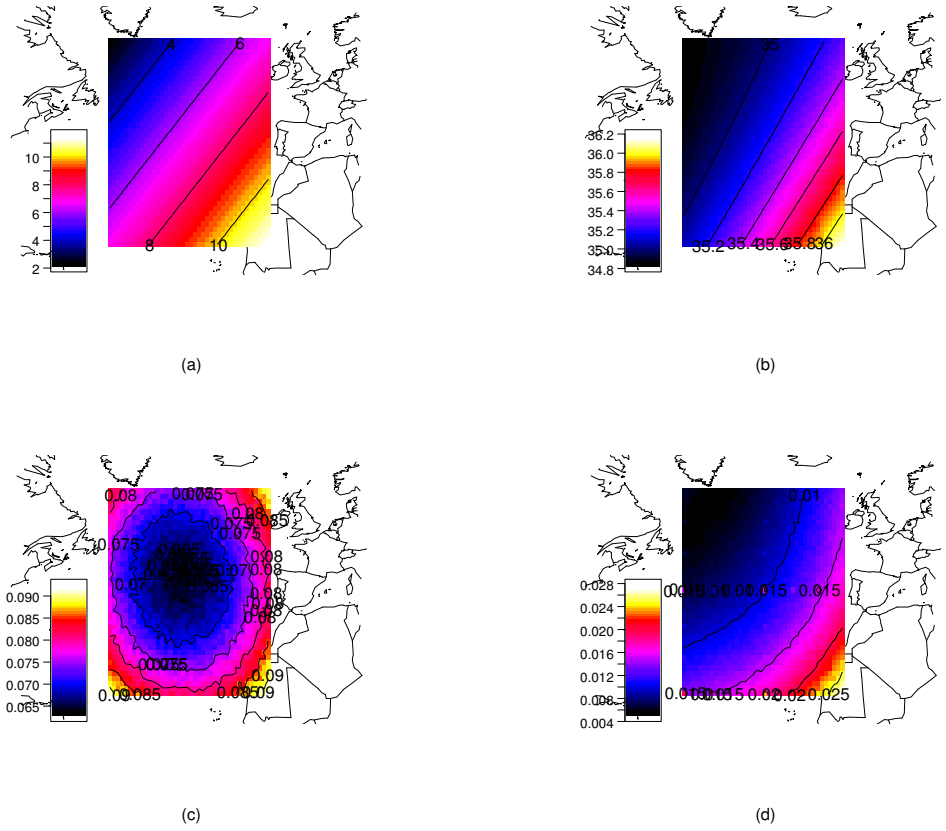

Figure 7: Annual prediction maps and their standard deviations at the deep ocean: temperature on panel (a); salinity on panel (b); standard deviation of temperature on panel (c); standard deviation of salinity on panel (d).