RICCI Sophie
Global Change and Climate Modeling Team
CERFACS - Toulouse, FRANCE

Technical Report TR/CMGC/01-88

Multivariate formulation of the background error covariance matrix

Contents

Forcing ocean model with windstress, heat flux and precipitation minus evaporation fields provides initial conditions from which to begin seasonal forecasts with coupled ocean atmospheric model. Temperature assimilation is currently used to improve the analysed ocean state and then turns out to be beneficial for seasonal climate forecasting. Temperature data assimilation in a global ocean model is currently used to initialize seasonal climate forecasts. The variational assimilation approach provides an increment vector, which added to the "background"or "previous forecast" state, gives the analysed state vector of the system.The spatial and multivariate structure of the analysis increment is mainly formulated in the background error covariance matrix, especially for a three dimensional variational assimilation system.

The background error covariance matrix B should express the true forecast errors in the assimilation, this formulation would imply the knowledge of the "true state" vector. Since there is no way to know this state, nor to calculate B given it's big size, the background error covariance matrix must be modeled.

Derber and Bouttier (1998) [3]propose a formulation of B as an operator approximating the variances, auto correlations, and cross correlations of the background errors. This operator isolates the multivariate aspect of the assimilation system in a linear balance operator K transforming unbalanced variables to the model variables.

The univariate aspect is divided in two operators: a diagonal Σ of background standard error deviation and a correlation operator C defining spatial correlation function for each field of the state vector.

The short-range errors in the assimilation are expressed by Weaver and Courtier (2001) [8] as correlation functions on the sphere for each field of the state vector. This method relies on a generalized horizontal diffusion equation, introducing the diffusivity coefficient κ. The solution of this generalized diffusion equation defines a correlation function in the spherical coordinates system. In the case of a simple horizontal diffusion equation, the correlation function can be approximated to a Gaussian of length scale $2\kappa T$. The vertical correlations are defined as the solution of a one dimensional diffusion equation. The correlation scales can be chosen anisotropic to model zonal stretching near the equator for example, and the "horizontal" correlation can be computed over isopycnal surfaces. Using this decomposition of the background error covariance matrix leads to a flow dependency in the spatial distribution of the information.

With a univariate assimilation system, the analysis only applies to the assimilated variable and may lead to inconsistent increment. Troccoli and Haines [2] underlined the fact that if temperature is corrected with assimilation while salinity is left unchanged, unrealistic water mass could be created and corrupt the simulated ocean state. In order to avoid this artificial mixing of the water column, the approach described in Troccoli 2000 ([1]) is to assimilate temperature, fit the analysed temperature profile to the background temperature profile, and then correct salinity using the $T - S$ relationship for the background. The salinity increment is then calculated "a posteriori", locally with respect to the background.

The solution adopted here includes the $T-S$ relationship in the background covariance matrix itself working with a decomposition in balanced and unbalanced parts of the increment vector. The salinity increment (and also the other fields increment) is then "a priori" calculated localy with respect to the background during the assimilation cycle.

The balanced components of the increment are linked to the unbalanced components and to a total variable(T in the following developments) through the operator K' . Given the increment in temperature and inversing the balance operator named K, the increment for all the variables is defined. This way, the analysis applies to all the variables of the state vector. The balance operator K guarantees the coherence of the increment vector and appears like a general alternative to Troccoli and Haines ([2]) method.

The article is organized as follows. Section 1 briefly presents the variational assimilation methods 3DVAR and 4DVAR and the formulation of the background error covariance matrix as a sequence of operators. The standard deviation matrix Σ, the correlation operator C, and the balance operator K are presented. Sections 2 and 3 aim at explaining the construction of the univariate members of the decomposition of B for the unbalanced components of the state vector. The multivariate relationships of the balance operator K are described in section 4, with a development of the $T - S$ preservation hypothesis. Finally, section 5 presents the ensemble of runs used to statistically determine some elements of the linear system K.

1 Formulation of the background error covariance matrix

1.1 Variational assimilation

The variational methods for data assimilation propose the minimization of a cost function J that estimates the difference between the simulated state of the system, the reference state and the observed state. The reference state is often chosen as the state of the system from the previous forecast also called "background" or "reference". The result of the assimilation consists in an increment added to the reference state in order to obtain the analysed state of the system. The control vector of the analysis is the increment vector $\delta\mathbf{x}$.

The cost functions for 3 dimensional and 4 dimensional incremental assimilation methods are respectively:

$$\begin{aligned} J(\delta\mathbf{x}) &= J_b(\delta x) + J_o(\delta\mathbf{x}) \\ &= \frac{1}{2}\delta\mathbf{x}^{\mathbf{T}}\mathbf{B}^{-1}\delta\mathbf{x} + \frac{\mathbf{1}}{\mathbf{2}}(\mathbf{H}\delta\mathbf{x} - \mathbf{d})^{\mathbf{T}}\mathbf{R}^{-1}(\mathbf{H}\delta\mathbf{x} - \mathbf{d}) \end{aligned} \tag{1}$$

$$J(\delta\mathbf{x}) = \frac{1}{2}\delta\mathbf{x}^{\mathbf{T}}\mathbf{B}^{-1}\delta\mathbf{x} + \frac{\mathbf{1}}{\mathbf{2}}\sum_{\mathbf{i=1}}^{\mathbf{N}}(\mathbf{H}_{\mathbf{x_{ti}}}\delta\mathbf{x_{ti}} - \mathbf{d_{ti}})^{\mathbf{T}}\mathbf{R_{ti}}^{-1}(\mathbf{H}_{\mathbf{x_{ti}}}\delta\mathbf{x_{ti}} - \mathbf{d_{ti}}) \tag{2}$$

where the increment state vector $\delta\mathbf{x}$ is the difference between the background estimation $\mathbf{x_b}$ and the simulated state $\mathbf{x}$.

$$\delta\mathbf{x} = \mathbf{x} - \mathbf{x_b} \tag{3}$$

In the case of a 4D Var, the increment is propagated in time by the tangent linear model M.

$$\delta\mathbf{x_{ti}} = \mathbf{M}_{\mathbf{t_i t_{i-1}}}\delta\mathbf{x}_{\mathbf{t_{i-1}}} \tag{4}$$

The minimization of the cost function gives the increment $\delta\mathbf{x}$ that optimizes the value of the analyzed state $\mathbf{x}$, given the forecast $\mathbf{x_b}$ and the observation vector $\mathbf{y^o}$ and their associated weighting error covariance matrix $\mathbf{B}$ and $\mathbf{R}$.

The new information given by the observation and the forecast is introduced in the minimization function thanks to the innovation vectors $\mathbf{d}$ or "departures". The innovation vectors $\mathbf{d}$ represent the difference between the observation vector $\mathbf{y^o}$ and $\mathbf{y}$, vector of the observation space, associated to $\mathbf{x_b}$ through the observation operator $\mathbf{H}$.

J_o (1) measures the difference between the innovation vector and the model increment, and $J_b(1)$ measures the fit to the background state.

In the described case, the assumption is made that the model is perfect, if not, J must include the following term::$\frac{1}{2}\epsilon^T\mathbf{R}^{-1}\epsilon$ where $\mathbf{Q}$ is the model error covariance matrix. Our study will focus on the strong constraint problem where $\mathbf{Q}$ is null.

A better preconditionned minimization problem is expressed as a function of the new variable $\mathbf{v}$, given by:

$$\mathbf{v} = \mathbf{B}^{\frac{-1}{2}}\delta\mathbf{x} \tag{5}$$

where $\mathbf{B}^{\frac{1}{2}}$ is taken to be any “square root” matrix such that $\mathbf{B} = \mathbf{B}^{\frac{T}{2}}\mathbf{B}^{\frac{1}{2}}$. The cost function for the 3D Var and 4D Var respectively, then can be written as functions of $\mathbf{v}$:

$$\begin{aligned} J(\mathbf{v}) &= \frac{1}{2}\mathbf{v}^{\mathbf{T}}\mathbf{v} + \frac{\mathbf{1}}{\mathbf{2}}(\mathbf{H}\delta\mathbf{x} - \mathbf{d})^{\mathbf{T}}\mathbf{R}^{-\mathbf{1}}(\mathbf{H}\delta\mathbf{x} - \mathbf{d}) && (6) \\ J(\delta\mathbf{x}) &= \frac{1}{2}\mathbf{v}^{\mathbf{T}}\mathbf{v} + \frac{\mathbf{1}}{\mathbf{2}}\sum_{\mathbf{i=1}}^{\mathbf{N}}(\mathbf{H}_{\mathbf{x_{ti}}}\delta\mathbf{x_{ti}} - \mathbf{d_{ti}})^{\mathbf{T}}\mathbf{R_{ti}}^{-\mathbf{1}}(\mathbf{H}_{\mathbf{x_{ti}}}\delta\mathbf{x_{ti}} - \mathbf{d_{ti}}) \end{aligned}$$

$$(7)$$

with $\delta\mathbf{x} = \mathbf{B}^{\frac{1}{2}}\mathbf{v}$.

To calculate the cost function is calculated the direct model is integrated on the assimilation window. The minimun of J is reached using a descent method algorithm, such as M1QN3. The minimization starts with $\mathbf{v} = \mathbf{0}$. Each iteration implies the computation of the cost function and it's gradient with respect to $\mathbf{v}$:

$$\begin{gathered} \nabla_v J = \nabla_v J_b + \nabla_v J_o \\ \nabla_v J = \mathbf{v} + \mathbf{B}^{\frac{T}{2}}\nabla_{\delta\mathbf{x}} J_o \end{gathered} \tag{8}$$

with $\nabla_{\delta\mathbf{x}} J_o = \mathbf{H}^{\mathbf{T}}\mathbf{R}^{-\mathbf{1}}(\mathbf{H}\delta\mathbf{x} - \mathbf{d})$.

For a 3D VAR method, the gradient of the cost function is directly calculated for each observation time.

$$\nabla_{\delta\mathbf{x}} J = \mathbf{B}^{-\mathbf{1}}\delta\mathbf{x} + \mathbf{H}^{\mathbf{T}}\mathbf{R}^{-\mathbf{1}}(\mathbf{H}\delta\mathbf{x} - \mathbf{d}) \tag{9}$$

A 4D VAR assimilation method, implies the backward integration of the adjoint of the tangent linear model to find the value of the gradient of J with respect to $\mathbf{v}$. The gradient of the cost function with respect to $\mathbf{v}$ is computed over the entire assimilation period for each iteration in the minimization inner loop.

$$\nabla_{\delta \mathbf{x}} J = \delta \mathbf{x}_{\mathbf{0}}^{*} \tag{10}$$

where x^* is the adjoint variable.

The formulation of $\mathbf{B}$ is crucial to the performance of the analysis system. Let's analyse the situation of a single assimilated observation of the value of a model field at one grid point, first with a 3D VAR and then with a 4D VAR. Since the gradient of the cost function must be null at the end of the minimization:

$$J_{\delta x} = \frac{1}{2}\delta \mathbf{x}^{\mathbf{T}}\mathbf{B}^{-1}\delta \mathbf{x} + \frac{\mathbf{1}}{\mathbf{2}}(\mathbf{H}\delta \mathbf{x} - \mathbf{d})^{\mathbf{T}}\mathbf{R}^{-1}(\mathbf{H}\delta \mathbf{x} - \mathbf{d}) \tag{11}$$

$$\begin{aligned} \nabla_{\delta x} J &= \mathbf{B}^{-1}\delta \mathbf{x} + \mathbf{H}^{\mathbf{T}}\mathbf{R}^{-1}(\mathbf{H}\delta \mathbf{x} - \mathbf{d}) \\ &= 0 \end{aligned} \tag{12}$$

Leading to :

$$\delta \mathbf{x} = \mathbf{B}\mathbf{H}^{\mathbf{T}}\mathbf{R}^{-1}(\mathbf{d} - \mathbf{H}\delta \mathbf{x}) \tag{13}$$

This formulation is equivalent to:

$$\begin{aligned} \delta \mathbf{x} &= -\mathbf{B}\mathbf{H}^{\mathbf{T}}\mathbf{R}^{-1}(\mathbf{H}(\mathbf{x} - \mathbf{x_b}) - (\mathbf{y^0} - \mathbf{H}\mathbf{x_b})) \\ &= -\mathbf{B}\mathbf{H}^{\mathbf{T}}\mathbf{R}^{-1}(\mathbf{H}\mathbf{x} - \mathbf{H}\mathbf{x_b}) - \mathbf{y^0} + \mathbf{H}\mathbf{x_b}) \\ &= \mathbf{B}\mathbf{H}^{\mathbf{T}}\mathbf{R}^{-1}(\mathbf{y^0} - \mathbf{H}\mathbf{x}) \end{aligned} \tag{14}$$

For this simple case:

$$\mathbf{y^0} - \mathbf{H}\mathbf{x} = (\mathbf{0.0.0...y} - \mathbf{x_k}\mathbf{...0.0.0...})^{\mathbf{T}} \tag{15}$$

The increment is then given by:

$$\delta \mathbf{x} = \mathbf{B}(\mathbf{0.0.0...}\frac{(\mathbf{y} - \mathbf{x_k})}{\sigma_{\mathbf{0}}^{\mathbf{2}}}\mathbf{...0.0.0...})^{\mathbf{T}} \tag{16}$$

That is, $\delta \mathbf{x}$ is proportional to a column of $\mathbf{B}$.

In the case of the 4D VAR the increment is given by :

$$\delta\mathbf{x} = (\mathbf{MBM^T})\mathbf{H^T R^{-1}}(\mathbf{d_k} - \mathbf{HM}\delta\mathbf{x}(\mathbf{t_0})) \tag{17}$$

where $\mathbf{H^T R^{-1}}(\mathbf{d_k} - \mathbf{HM}\delta\mathbf{x}(\mathbf{t_0}))$ is a vector of zeros except at the observation point. The increment is, in this case, proportional to a column of $\mathbf{MBM^T}$.

In a 4D VAR, $\mathbf{B}$ is implicitely propagated in time by the model $\mathbf{M}$:

$$\begin{aligned} E[(\delta\mathbf{x}(\mathbf{t_0}))(\delta\mathbf{x}(\mathbf{t_0}))^\mathbf{T}] &= \mathbf{B} \\ E[(\delta\mathbf{x}(\mathbf{t_k}))(\delta\mathbf{x}(\mathbf{t_k}))^\mathbf{T}] &= [\mathbf{M}\delta\mathbf{x}(\mathbf{t_0})][\mathbf{M}\delta\mathbf{x}(\mathbf{t_0})]^\mathbf{T} \\ &= \mathbf{M}[\delta\mathbf{x}(\mathbf{t_0})][\delta\mathbf{x}(\mathbf{t_0})]^\mathbf{T}\mathbf{M^T} \\ &= \mathbf{MBM^T} \end{aligned} \tag{18}$$

1.2 Formulation of the multivariate background error covariance matrix B

The matrix $\mathbf{B}$ and $\mathbf{R}$ are respectively the background error covariance matrix and the observation error covariance matrix.

The background errors $\mathbf{e}$ are defined as the difference between the forecast state vector (or background) $\mathbf{x}_b$ and the true value of the state vector $\mathbf{x_t}$. The observations errors $\mathbf{e}_o$ are defined as the difference between the observation vector $\mathbf{y}^o$ and $\mathbf{y}$, vector of the observation space, associated to $\mathbf{x}_t$ through the observation operator $\mathbf{H}$.

$$\mathbf{e}_b = \mathbf{x_b} - \mathbf{x_t} \tag{19}$$

$$\mathbf{e}_o = \mathbf{y_o} - \mathbf{H}(\mathbf{x_t}) \tag{20}$$

and the error covariance matrix are formulated as follows:

$$\begin{aligned} \mathbf{B} &= E[(\mathbf{e}^b - E(\mathbf{e}^b)) * (\mathbf{e}^b - E(\mathbf{e}^b))^T] \\ \mathbf{B} &= E[(\mathbf{e}^b)(\mathbf{e^b})^\mathbf{T}] \end{aligned} \tag{21}$$

$$\begin{aligned} \mathbf{R} &= E[(\mathbf{e}^o - E(\mathbf{e}^o)) * (\mathbf{e}^o - E(\mathbf{e}^o))^T] \\ \mathbf{R} &= E[(\mathbf{e}^o)(\mathbf{e^o})^\mathbf{T}] \end{aligned} \tag{22}$$

where E[.] denotes mathematical expectation.

It follows that these two covariance matrix are symetrical, and positive defined.

Specifying $\mathbf{B}$, we want to describe the statistics of the errors in the background. However we don't know what the errors in the background are, since we don't know the true state of the ocean. The other problem to be addressed in specifying $\mathbf{B}$ is that this matrix is enormous and even if we could fit it into a computer, we don't have enough statistical information to determine all it's elements. As a consequence, $\mathbf{B}$ must be modeled. Since the information in $\mathbf{B}$ has a strong influence on the analysis (see section 1.1), it's computation should though be as close as possible to the true background error covariance matrix, especially speaking in terms of relations between the model variables, and spatial correlation for each variable.

The diagonal elements of $\mathbf{B}$ correspond to the standard error deviation for each model variable. The off-diagonal elements correspond to covariances. The block-diagonal elements are the correlations between grid points for a particular model variable, it describes the spreading of the information on the grid and should provide consistent increments at the neighbourhood grid points and levels of the model.

The other elements are the covariances between two different model variables on the grid. These last covariances describe the distribution of the assimilated information for one field upon all the other fields of the state vector, this gives a the multivariate aspect to the analysis, and should ensure that observations of one model variable produce dynamically consistent increments for the other model variables.

Derber and Bouttier (1998) [3] have proposed a separation of $\mathbf{B}$ in several operators in order to isolate the univariate from the multivariate aspects of the error covariance matrix. This formulation imposes the partitioning of the model variables into balanced and unbalanced components, except for one variable which is taken in totality.

The increment is presented as the sum of a balanced vector $\delta\mathbf{x_B}$ and an unbalanced vector $\delta\mathbf{x_U}$:

$$\begin{aligned}\delta\mathbf{x} &= \delta\mathbf{x}_B + \delta\mathbf{x}_U \\ &= \mathbf{K}^{'}\delta\mathbf{x_U} + \delta\mathbf{x_U} \\ &= (\mathbf{K}^{'} + \mathbf{I})\delta\mathbf{x_U} \\ &= \mathbf{K}\delta\mathbf{x_U} \end{aligned} \tag{23}$$

The total variable is used to establish balances with the other variables through the linear system $K = K'+I$ that converts the unbalanced components to the balanced components of the increment variables.

The univariate operators Σ and C apply to the unbalanced components of the increment vector and the balance operator K leads to the full computation of the analysed increment.

The background error covariance matrix $\mathbf{B}$ is expressed as the operator:

$$\mathbf{B} = \mathbf{K}\mathbf{\Sigma_U}\mathbf{C_U}\mathbf{\Sigma_U}\mathbf{K^T} \tag{24}$$

and the preconditionned increment $\mathbf{v}$ is

$$\begin{aligned} \delta\mathbf{x} &= \mathbf{B}^{\frac{1}{2}}\mathbf{v} \\ &= \mathbf{K}\mathbf{\Sigma_U}\mathbf{C_U^{\frac{1}{2}}}\mathbf{v} \end{aligned} \tag{25}$$

where:

- $\mathbf{\Sigma_U}$ is a diagonal matrix of background error standard deviation for the unbalanced components of the model variables

- $\mathbf{C_U}$ is a block diagonal symetrical matrix of background error correlations for the unbalanced components of the model variables. Each block contains the spatial correlation between different position on the grid for a single variable.

As a consequence the matrix $\mathbf{B_U} = \mathbf{\Sigma_U}\mathbf{C_U}\mathbf{\Sigma_U}$ has a block diagonal structure with no correlation between pairs of different model variables. The multivariate notion, is fully contained in the matrix K.

The presented method is applied on the ocean model OPA, where the state vector of prognostic variables is (T, S, U, V) in the case of a rigid-lid, and (T, S, U, V, η) in the case of a free surface.

In this last case, and if T is chosen as the total variable (this assumption will be developed in part (2), the matrix $\mathbf{B_U}$ is:

$$B_U = \begin{pmatrix} Cov(T) & 0 & 0 & 0 & 0 \\ 0 & Cov(S_U) & 0 & 0 & 0 \\ 0 & 0 & Cov(U_U) & 0 & 0 \\ 0 & 0 & 0 & Cov(V_U) & 0 \\ 0 & 0 & 0 & 0 & Cov(\eta_U) \end{pmatrix} \tag{26}$$

The modelisation of the background error covariance matrix goes through the definition of the operators $\mathbf{\Sigma}$, $\mathbf{C}$ with respect to the unbalanced components of the state vector variables and the balance operator $\mathbf{K}$. For a 3D variational method, these matrix won't evolve in time, whereas, for a 4D variational method, these matrix will implicitly be propagated in time over the data assimilation period by the integration of the direct model as exposed in part 1.1.

2 The background variance matrix Σ^2 for the unbalanced components of the increment vector

2.1 How to estimate the variances

In a multivariate formulation of variational data assimilation, the matrix $\mathbf{\Sigma_U}$ and $\mathbf{C_U}$ described in the first part of this paper apply to the unbalanced components of the increment vector, whereas in a univariate case they would apply to the total vector increment.

Here, the elements in $\mathbf{\Sigma_U}$ must be the background standard error deviation for each unbalanced variable. These standard error deviation can be empirically defined with a climatology or estimated with statistics on an ensemble method.

For the moment, some statistics given as the standard error deviation for the total variables are still used, the computation of some statistics on the unbalanced part of the variables on an ensemble of runs is one of our next task.

Our approach of the ensemble method is the following: The first step is to run the model with a given wind and sea surface temperature forcing, and keep this as a "reference run". Then, the wind field is successively modified with slightly different pertubations, and the model is run with each of these perturb wind forcing fields.

The NMC ensemble approach described by Gauthier and Buehner ([7]), also generates an ensemble of states, but the runs are computed with different initial conditions. A forecast with assimilation is started at time t_0, and 24 hours later, a second forecast is started with the state of the system at time t_{24} as initial condition, taking into account the information added with the assimilation during the first 24 hours. Twenty four hours later, at time t_{48}, 2 forecasts are generated, with a sizeable dispersion.

In a univariate case, one would calculate the variances straight from the total variables of this ensemble of runs.

2.2 The background variances for unbalanced variables

As presented in (13), the total state vector is formulated as the sum of balanced and unbalanced components $\delta\mathbf{x_B}$ and $\delta\mathbf{x_U}$.

In Derber and Boutier (1998) [3] formulation, the balanced components of the state vector are linearly dependent on a single total variable (T in our case), and on the unbalanced components of the state vector. Given a vector δT and the linear relationships between balanced and unbalanced components, the inversion of the system (23) determines the unbalanced components of the vector. Therefore, from an ensemble of vectors $\delta \mathbf{x}$, one can deduce the corresponding ensemble of vectors $\delta \mathbf{x_U}$ and then compute the standard deviation statistics.

To sum up this part,the different steps to compute the operator $\mathbf{\Sigma}$ are as follows:

- First, generate an ensemble of runs, here with different wind forcing (this part is developed in part 5),

- Define the balance relationships of the linear system $\mathbf{K}$

- Compute $\delta \mathbf{x_U}$ and calculate the variances for each unbalanced component of the model variables.

3 The background error correlation matrix C for the unbalanced components of the increment vector

3.1 Definition

The formulation of the univariate correlation operator $\mathbf{C}$ as an operator make it well adapted for large state vector assimilation problems. As discussed in Weaver and Courtier 2001([8], this operator can be modeled as :

$$\mathbf{C_U} = \mathbf{\Lambda L^{\frac{T}{2}} W^{-1} L^{\frac{1}{2}} \Lambda} \tag{27}$$

The role of the operators Λ, L, W will be exposed in the following sections.

Since the elements in $\mathbf{C}$ are the background correlations for each unbalanced variable, the structure of $\mathbf{C_U}$ is block diagonal. The correlation operator $\mathbf{C}$ can then be split in individual operators $\mathbf{C}_\alpha$ acting on the different components v_α of the vector $\mathbf{v}$.

3.2 Modelling the correlation functions

The spreading of the information from the assimilated point to the entire grid is interpreted as a diffusion phenomenon. The spatial correlations are then described using a diffusion equation. The resolution of this partial derivative equation leads to a quasi-gaussian class of correlation functions (see section 3.3.2 - The Gaussian approximation), this ensemble can be extended using a generalized diffusion equation.

The horizontal correlations are modeled using the generalized diffusion equation on the sphere (28) where κ_p are positive weighting coefficients. In the simple case of diffusion equation, $\kappa_p = 0$ for all p>1.

$$\frac{\partial \eta}{\partial t} + \sum_{p=1}^{P} \kappa_p (-\nabla^2)^p \eta = 0 \tag{28}$$

The solution of (28) on the sphere where λ is longitude and ϕ is latitude is of the form :

$$\eta(\lambda, \phi, T) = \frac{1}{4\pi a^2} \int_{\Sigma'} f(\theta, \kappa_1 T, \kappa_2 T, ..., \kappa_p T) * \eta(\lambda', \phi', O) d\Sigma' \tag{29}$$

where θ is the angular separation between points on the sphere of radius a.

The integral kernel given by the functions

$$f(\theta, \kappa_1 T, \kappa_2 T, ..., \kappa_p T) = \sum_{n=0}^{\infty} \sqrt{2n+1} e^{-\sum_{p=1}^{P} \kappa_p T \frac{n(n+1)}{a^2}} P_n^o(cos(\theta) \tag{30}$$

defines a family of isotropic covariance function on the sphere. The length scale of correlation is defined by:

$$L^2 = -2 \frac{f(0, \kappa_1 T, ..., \kappa_p T)}{\delta^2 f(0, \kappa_1 T, ..., \kappa_p T)} \tag{31}$$

The univariate correlation matrix $\mathbf{C}_\alpha$ must account for error correlation in the vertical as well as in the horizontal. As discussed in Weaver and Courtier 2001([8]), the vertical correlation model is separated from the horizontal correlation model and defined as a one dimension diffusion equation of the form (23) where $\mathbf{K}$ is the diffusion coefficient.

$$\frac{\delta \eta}{\delta t} - \kappa \frac{\delta^2 \eta}{\delta z^2} = 0 \tag{32}$$

The solution is given by the convolution of the initial condition $\eta(z, 0)$ with a Gaussian function :

$$\eta(z, T) = \frac{1}{\sqrt{4\pi\kappa T}} \int_{z'} e^{-\frac{(z-z')^2}{4\kappa T}} \eta(z', O) dz' \tag{33}$$

where the length scale of the gaussian L is $\sqrt{2\kappa T}$. This integral equation,once normalized by Λ defines a covariance function on $\mathbb{R}$.

3.3 The Gaussian approximation and Implementation of a three-dimensional correlation model

3.3.1 Discrete representation of the generalized diffusion equation

The equation (28) is discretized with an explicit forward differencing scheme of the temporal derivative. The discretized Laplacian is represented by the operator D.

The integration of (28) over M time steps is represented by a discrete operator L_p defined as:

$$L_P = [I - \sum_{p=1}^{P} \kappa_p \Delta t (-D)^p]^M \tag{34}$$

The integral operator is equivalent to $\eta(t_M) = L_P \eta(t_0)$.

L_P can be factored as :

$$L_P = L_P^{\frac{1}{2}} W^{-1} L_P^{\frac{T}{2}} W \tag{35}$$

where W is a diagonal matrix of local area elements and

$L_P = L_P^{\frac{1}{2}} L_P^{\frac{1}{2}}$ with $L_P^{\frac{T}{2}} = (L_P^{\frac{1}{2}})^T$

The correlation matrix associated with the field α has the form

$$\mathbf{C}_\alpha = \mathbf{\Lambda} \mathbf{L}_\mathbf{P}^{\frac{1}{2}} \mathbf{W}^{-1} \mathbf{L}_\mathbf{P}^{\frac{T}{2}} \mathbf{\Lambda} \tag{36}$$

where $\mathbf{\Lambda}$ is a diagonal normalization matrix that converts the covariance matrix $L_P \mathbf{W}^{-1}$ to a correlation matrix $\mathbf{C}_\alpha$, where the diagonal elements are equal to one.

3.3.2 The Gaussian approximation

In order to reduce the cost and the complexity of the implementation, the horizontal correlation is taken as a quasi-Gaussian correlation function.

In other words, the partial differential equation (28)is reduced to a classical 2D diffusion equation :

$$\frac{\delta \eta}{\delta t} - \kappa \nabla^2 \eta = 0 \tag{37}$$

(P is equal to 1 in equation (28)).

The solution on the sphere still takes the form :

$$\eta(\lambda, \phi, T) = \frac{1}{4\pi a^2} \int_{\Sigma'} \int f(\theta, kT) \eta(\lambda', \phi', 0) d\Sigma' \tag{38}$$

A solution to (37) may be expanded as

$$\eta(\lambda, \phi, T) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \eta_m^n(t) Y_n^m(\lambda, \phi) \tag{39}$$

where $Y_n^m(\lambda, \phi)$ are the spherical harmonics.

$Y_n^m(\lambda, \phi)$ are the eigenvectors of the Laplacian operator on the sphere, with $\frac{-n(n+1)}{a^2}$ the associated eigenvalues, the diffusion equation turns into:

$$\frac{d\eta_m^n}{dt} = -\kappa \frac{-n(n+1)}{a^2} \eta_m^n \tag{40}$$

The solution of (40) is

$$\eta_m^n(T) = \eta_m^n(0)e^{-KT\frac{-n(n+1)}{a^2}} \tag{41}$$

with

$$\eta_m^n(0) = \frac{1}{4\pi a^2}\int_\Sigma \eta(\lambda,\phi,0)Y_n^{m^*}(\lambda,\phi)d\Sigma' \tag{42}$$

Replacing (42) in (41), and (41) in (39), the function $f(\theta,\kappa T)$ in (38) is given by:

$$f(\theta,\kappa T) = \sum_{n=0}^{\infty}(\sqrt{2n+1}e^{-\kappa T\frac{n(n+1)}{a^2}}\sum_{m=-n}^{n}Y_n^m(\lambda,\phi)Y_n^{m^*}\lambda,\phi) \tag{43}$$

which, applying the Addition Theorem for the spherical harmonics (Arfken 1966) turns into

$$f(\theta,kT) = \sum_{n=0}^{\infty}(\sqrt{2n+1}e^{-\kappa T\frac{n(n+1)}{a^2}}P_n^0(cos\theta) = \sum_{n=0}^{\infty}A_nP_n^0(cos\theta) \tag{44}$$

f is a correlation function by virtue of the positivity of A_n, and eventhough it is not a Gaussian, it is closely to match a Gaussian for small length scale compared to the radius of the earth(See Weaver and Courtier 2001 Appendix ([8])).

By modelling the spatial correlations with a diffusion equation (P = 1), we have fixed the shape of the correlation function to a Gaussian (or quasi-Gaussian). It's length scale is given by $2\kappa T$. This product is chosen in order to fit the Gaussian model to the ensemble statistics.

The estimation of f leads to the estimation of the parameter κT necessary to compute L_1 (see equation (34)), where T is the total integration time $T = M * \delta t$.

$$L_1 = [I - \kappa\delta t(-D)]^M \tag{45}$$

3.3.3 The horizontal and vertical separation

As explained in Weaver and Bouttier (2001) [8] three dimensional correlation models are separable in a product of a horizontal and a vertical correlation function.

The horizontal correlation function is taken as described in 3.3.2, as a solution of 2D diffusion equation, and the vertical correlation is treated as the solution of the 1 D diffusion equation (37).

Letting D_v denote the matrix representation of the discretized vertical operator $\frac{(\delta\eta)^2}{\delta z^2}$, and L_p^v the operator that integrates the 1 D diffusion equation (29) L_R^v from $t = 0$ to $t = T$.

$$L_R^v = [I - \kappa_r^v \delta t(-D^v)]_v^M \tag{46}$$

From now, we will use the subscripts h and v to distinguish the horizontal and vertical operators.

The three dimensional covariance operator is defined as follows:

$$L_r^v W_v^{-1} L_p^h W_h^- 1 = L_r^{v\frac{1}{2}} L_p^{h\frac{1}{2}} W^{-1} L_p^{h\frac{T}{2}} L_r^{v\frac{T}{2}} \tag{47}$$

where $W = W_h W_r$ is a diagonal matrix of volume elements.

Finally, the square-root factors of the three dimensional correlation matrix take the form:

$$C_\alpha^{\frac{T}{2}} = \Lambda L_r^{v\frac{1}{2}} L_p^{h\frac{1}{2}} W^{-\frac{1}{2}} C_\alpha^{\frac{1}{2}} = W^{-\frac{1}{2}} L_p^{h\frac{T}{2}} L_r^{v\frac{T}{2}} \Lambda \tag{48}$$

3.3.4 Anisotropic correlations and the choice of the coordinate system

Until now, we have made two important assumptions. On the one hand, the correlation functions were computed over horizontal and vertical surfaces, and on the other hand we assumed that the correlation scales were isotropic.Weaver and Courtier (2001) [8] propose an alternative to these “a priori” choices.

The anisotropy of correlation scales can be modeled by stretching the zonal coordinates in the Laplacian operator. The regular Laplacian operator is replaced by the operator $\nabla \cdot R\nabla$, R being a diagonal, second-rank tensor of non dimensional “diffusion” coefficients r_i, r_j:

$$R = \begin{pmatrix} r_i & 0 \\ 0 & r_j \end{pmatrix}$$

This stretching is especially useful near the equator where correlation scales are typically longer in the zonal direction then in the meridional direction.

The horizontal correlations are computed over the isopycnal surfaces in order to introduce a flow dependency into the background error correlation model.

This implies a reinterpretation of the tensor R in our new coordinate system, assuming that r_i and r_j are diffusion coefficient in the "horizontal plane" defined by the isopycnal surfaces, and r_k is the diffusion coefficient in the "vertical direction" normal to this surfaces.

The different tasks necessary to build the correlation operator $\mathbf{C}$ are as follows:

- Decide shape of correlation function to work on

- Estimate the length scale of the Gaussian correlation function doing statistics on a ensemble of runs.

- Give an explicit form of the operator L_1 (through the parameter κT)

- Precise the anisotropy constraints.

4 The multivariate relationships and the linear operator K

4.1 The choice of the total variable

The formulation of the multivariate relationships is mainly conditioned by the choice of the total variable. The choice of the total variable should be determined by the existence of dominant physical relationships between this variable and the other model variables. A reasonable knowledge of the error statistics of the total variable is necessary since this variable largely establishes the error statistics of the balanced part of the other variables. The most relevant option is to consider T as the total variable, since it is a 3 dimensional field and then gives information about barotropic as well as baroclinic movements. Moreover, T is the variable to assimilate and consequently the variable for which the error statistics is best known.

4.2 Formulation of the multivariate relationships

The increment vector $\delta\mathbf{x} = (\delta T, \delta S, \delta U, \delta V, \delta\eta)$ is expressed as follows:

$$\left\{ \begin{array}{rcl} \delta T & = & \delta T_B \\ & = & \delta T_U \\ \delta S & = & \delta S_B + \delta S_U \\ & = & f_{ST}\delta T + \delta S_U \\ \delta\eta & = & \delta\eta_B + \delta\eta_U \\ & = & f_{\eta T}\delta T + f_{\eta S}\delta S + \delta\eta_U \\ & = & (f_{\eta T} + f_{\eta S} f_{ST})\delta T + f_{\eta S}\delta S_U + \delta\eta_U \\ \delta U & = & \delta U_B + \delta U_U \\ & = & f_{UT}\delta T + f_{US}\delta S + f_{U\eta}\delta\eta + \delta U_U \\ & = & (f_{UT} + f_{US} f_{ST} + f_{U\eta} f_{\eta S} + f_{U\eta} f_{\eta S} f_{ST})\delta T + (f_{US} + f_{U\eta} f_{\eta S})\delta S_U + f_{U\eta}\delta\eta_U + \delta U_U \\ \delta V & = & \delta V_B + \delta V_U \\ & = & f_{VT}\delta T + f_{VS}\delta S + f_{V\eta}\delta\eta + \delta V_U \\ & = & (f_{VT} + f_{VS} f_{ST} + f_{V\eta} f_{\eta S} + f_{V\eta} f_{\eta S} f_{ST})\delta T + (f_{VS} + f_{V\eta} f_{\eta S})\delta S_U + f_{V\eta}\delta\eta_U + \delta V_U \end{array} \right. \tag{49}$$

This linear system converts the unbalanced variables into complete variables expressing the balanced part of the increment as a linear combinaison of unbalanced components.

$$\left\{\begin{array}{rcl} \delta T &=& \delta T_U \\ \delta S &=& K_{ST}\delta T + \delta S_U \\ \delta\eta &=& K_{\eta T}\delta T + K_{\eta S}\delta S_U + \delta\eta_U \\ \delta U &=& K_{UT}\delta T + K_{US}\delta S_U + K_{U\eta}\delta\eta_U + \delta U_U \\ \delta V &=& K_{VT}\delta T + K_{VS}\delta S_U + K_{V\eta}\delta\eta_U + \delta V_U \end{array}\right. \tag{50}$$

The linear system $\mathbf{K}$ then becomes:

$$\mathbf{K} = \begin{pmatrix} I & 0 & 0 & 0 & 0 \\ K_{ST} & I & 0 & 0 & 0 \\ K_{\eta T} & K_{\eta S} & I & 0 & 0 \\ K_{UT} & K_{US} & K_{U\eta} & I & 0 \\ K_{VT} & K_{VS} & K_{V\eta} & 0 & I \end{pmatrix}$$

with:

$$\left\{\begin{array}{rcl} K_{ST} &=& f_{ST} \\ K_{\eta T} &=& f_{\eta T} + f_{\eta S}f_{ST} \\ K_{\eta S} &=& f_{\eta S} \\ K_{UT} &=& f_{UT} + f_{US}f_{ST} + f_{U\eta}f_{\eta S} + f_{U\eta}f_{\eta S}f_{ST} \\ K_{US} &=& f_{US} + f_{U\eta}f_{\eta S} \\ K_{U\eta} &=& f_{U\eta} \\ K_{VT} &=& f_{VT} + f_{VS}f_{ST} + f_{V\eta}f_{\eta S} + f_{V\eta}f_{\eta S}f_{ST} \\ K_{VS} &=& f_{VS} + f_{V\eta}f_{\eta S} \\ K_{V\eta} &=& f_{V\eta} \end{array}\right. \tag{51}$$

The analysis leads to the increment vector $\delta\mathbf{x}$ in equation (27) in section 1.1:

$$\left\{\begin{array}{rcl} \delta\mathbf{x} &=& \mathbf{BH^TR^{-1}}(\mathbf{d} - \mathbf{H}\delta\mathbf{x}) \\ &=& (\mathbf{B^{-1}} + \mathbf{H^TR^{-1}H})^{-1})\mathbf{H^TR^{-1}} \\ &=& \mathbf{BH^T}(\mathbf{HBH^T} + \mathbf{R^{-1}})\mathbf{d} \end{array}\right. \tag{52}$$

Let ω be the vector $\mathbf{H^T}(\mathbf{R} + \mathbf{HBH^T})^{-1}\mathbf{d}$, in the case the assimilated field is temperature, ω express as follows:

$$\omega = \begin{pmatrix} \omega_{\mathbf{T}} \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{53}$$

the increment is then proportional to the first column of B :

$$\delta\mathbf{x} = \begin{pmatrix} B_{T_U}\omega_{\mathbf{T}} \\ K_{ST}B_{T_U}\omega_{\mathbf{T}} \\ K_{\eta T}B_{T_U}\omega_{\mathbf{T}} \\ K_{UT}B_{T_U}\omega_{\mathbf{T}} \\ K_{VT}B_{T_U}\omega_{\mathbf{T}} \end{pmatrix} \tag{54}$$

As a consequence, this computation only requires the definition of $\mathbf{B_{T_U}}$, which is the block of the matrix B_U defined in 1.2 as $\mathbf{B_U} = \mathbf{\Sigma_U C_U \Sigma_U}$ related to temperature standard error deviations and correlations.

Gauthier and Buehner (1998) [7] have proposed a simplified formulation of the operator $\mathbf{K}'$ that establishes linear balances between the balanced components only.

The increment $(\delta T, \delta S, \delta U, \delta V, \delta\eta)$ is expressed as follows:

$$\left\{\begin{array}{rcl} \delta T & = & \delta T_B \\ & = & \delta T_U \\ \delta S & = & \delta S_B + \delta S_U \\ & = & f^1_{ST}\delta T + \delta S_U \\ \delta\eta & = & \delta\eta_B + \delta\eta_U \\ & = & f^1_{\eta T}\delta T + f^1_{\eta S}\delta S_B + \delta\eta_U \\ & = & (f^1_{\eta T} + f^1_{\eta S} f^1_{ST})\delta T + \delta\eta_U \\ \delta U & = & \delta U_B + \delta U_U \\ & = & f^1_{UT}\delta T + f^1_{US}\delta S_B + f^1_{U\eta}\delta\eta_B + \delta U_U \\ & = & (f^1_{UT} + f^1_{US} f^1_{ST} + f^1_{U\eta} f^1_{\eta S} + f^1_{U\eta} f^1_{\eta S} f^1_{ST})\delta T + \delta U_U \\ \delta V & = & \delta V_B + \delta V_U \\ & = & f^1_{VT}\delta T + f^1_{VS}\delta S + f^1_{V\eta}\delta\eta + \delta V_U \\ & = & (f^1_{VT} + f^1_{VS} f^1_{ST} + f^1_{V\eta} f^1_{\eta S} + f^1_{V\eta} f^1_{\eta S} f^1_{ST})\delta T + \delta V_U \end{array}\right. \tag{55}$$

This linear system converts the unbalanced variables into complete variables expressing the balanced part of the increment as a function of the total variable.

$$\left\{\begin{array}{rcl} \delta T & = & \delta T_U \\ \delta S & = & K_{ST}\delta T + \delta S_U \\ \delta\eta & = & K_{\eta T}\delta T + \delta\eta_U \\ \delta U & = & K_{UT}\delta T + \delta U_U \\ \delta V & = & K_{VT}\delta T + \delta V_U \end{array}\right. \tag{56}$$

The linear system $\mathbf{K}$ then becomes:

$$\mathbf{K} = \begin{pmatrix} I & 0 & 0 & 0 & 0 \\ K_{ST} & I & 0 & 0 & 0 \\ K_{\eta T} & 0 & I & 0 & 0 \\ K_{UT} & 0 & 0 & I & 0 \\ K_{VT} & 0 & 0 & 0 & I \end{pmatrix} \tag{57}$$

with the operators $\mathbf{K_{ij}}$ defined previously.

$$
\left\{
\begin{array}{lcl}
K_{ST} & = & f^1_{ST} \\
K_{\eta T} & = & f^1_{\eta T} + f^1_{\eta S} f^1_{ST} \\
K_{\eta S} & = & 0 \\
K_{UT} & = & f^1_{UT} + f^1_{US} f^1_{ST} + f^1_{U\eta} f^1_{\eta S} + f^1_{U\eta} f^1_{\eta S} f^1_{ST} \\
K_{US} & = & 0 \\
K_{U\eta} & = & 0 \\
K_{VT} & = & f^1_{VT} + f^1_{VS} f^1_{ST} + f^1_{V\eta} f^1_{\eta S} + f^1_{V\eta} f^1_{\eta S} f^1_{ST} \\
K_{VS} & = & 0 \\
K_{V\eta} & = & 0
\end{array}
\right. \tag{58}
$$

In this case, the balanced vector $\delta \mathbf{x_B}$ takes the simple form:

$$
\underline{\delta x_B} = K \ (\delta T \quad 0 \quad 0 \quad 0 \quad 0)^T \tag{59}
$$

The increment is again proportional to the first column of B :

$$
\delta \mathbf{x} = \begin{pmatrix} B_{T_U} \omega_{\mathbf{T}} \\ K_{ST} B_{T_U} \omega_{\mathbf{T}} \\ K_{\eta T} B_{T_U} \omega_{\mathbf{T}} \\ K_{UT} B_{T_U} \omega_{\mathbf{T}} \\ K_{VT} B_{T_U} \omega_{\mathbf{T}} \end{pmatrix} \tag{60}
$$

4.3 Determination of the balance relationships

The formulation of the multivariate relationships defined in (23) and (49) requires the determination of the linear functions f_{ij} presented in section 4.2.

One way to estimate these functions is to use an ensemble approach (Parrish and Derber (1992) [6], see also Gauthier and Buehner (1998) [7]).

The OPA model is forced with slightly different winds in order to generate an ensemble of results (E) (A description of these experiments will be done elsewhere). This ensemble is already used in section 2, to compute the error standard deviation in $\mathbf{\Sigma_U}$. It can also be used to determine some coefficients of the functions f_{ij}, for example using linear regressions. This possibility is illustrated in the following subsection with the variables $T - S$.

4.3.1 The $T - S$ preservation hypothesis

Since T is the variable kept total, the next step in the definition of the operator $\mathbf{K}$ is the construction of a linear relationship between temperature and salinity. Troccoli and Haines (1999) ([2]) underlines the major negative impact that temperature assimilation can have on the water mass column stability when the salinity is left unmodified. If the temperature increments given by the assimilation scheme are such to decrease the vertical T stratification while S is left unchanged, the water column may become unstable. Troccoli and Haines(1999) ([2]) shows the unrealistic temperature structure in the equatorial region three months after the assimilation of TAO measurements with no correction of the salinity fields. The main purpose is then to find an approach to include salinity field adjustment in a data assimilation context when only temperature measurements are available. The solution proposed by Troccoli and Haines is based on the $T-S$ preservation hypothesis. Vossepoel et al(1999) ([4]) found that a long-term mean $T-S$ relation based on some climatological dataset represents a good first guess field but that corrections to the salinity field are necessary to get a reliable density field. In order to avoid this "a posteriori" correction, Troccoli used a local salinity analysis, with the background $T-S$ properties being taken either from those most recently available from CTD data or from a model prediction of $T-S$ in the context of a sequential data assimilation.

The method described in this section uses a similar local relation $T-S$ taken from the model background, in order to build a linear relation between T and the balanced components of S. But here, the $T-S$ relation is directly included in the background error covariance matrix $\mathbf{B}$. Let δT be the increment in temperature, S can be linearized with respect to the background field as follows:

$$S(T) = S(T_b + \delta T) = S(T_b) + \alpha \frac{\delta S_b}{\delta T_b} \delta T \tag{61}$$

The balanced components of S will then be represented by

$$\begin{aligned} \delta S &= \delta S_B + \delta S_U \\ &= \alpha \tfrac{\delta S_b}{\delta T_b} \delta T + \delta S_U \end{aligned} \tag{62}$$

where α may be a function of position on the grid or of more "physical parameters such as stratification ρ_b, Brunt-Väisälä frequency N_b or Richardson Ri_b number with respect to the background, and also indirectly depends on the space and time coordinates.

The expression of $\alpha(\rho, N, ..)$ is determined with a statistical approach on the ensemble (E) with a linear regression.

The derivative of salinity with respect to temperature is calculated along the vertical in order to conserve the potential vorticity.

The same approach can be conserved to identify the density field in the linear equation of state:

$$\begin{aligned} \delta\rho &= \lambda 1 \delta T + \lambda 2 \delta S \\ &= \tfrac{\delta\rho_b}{\delta T_b}\delta T + \tfrac{\delta\rho_b}{\delta S_b}\delta S \\ &= \tfrac{\delta\rho_b}{\delta T_b}\delta T + \tfrac{\delta\rho_b}{\delta S_b}(\alpha\tfrac{\delta S_b}{\delta T_b}\delta T + \delta S_U) \end{aligned} \tag{63}$$

4.3.2 The geostrophic balance hypothesis and the sea surface height expression

The pressure for the water column is given by:

$$P = \int_z^0 (\delta\rho g z)dz + \rho_0 g\eta \tag{64}$$

where $\rho_0 g\eta$ is the surface pressure P_s.

Assuming that the hydrostatic hypothesis is valid, the geostrophic balance hypothesis can be expressed as (65) linking velocity and density:

$$\begin{aligned} \delta u_f &= -\tfrac{g}{f\rho_0}\tfrac{\delta P}{\delta y} \\ \delta v_f &= \tfrac{g}{f\rho_0}\tfrac{\delta P}{\delta x} \end{aligned} \tag{65}$$

where the subscript f goes for f-plane hypothesis.

At the equator where the geostrophy breaks down ($f->0$), another physical relation should be chosen (see Picaud 1990 (/citepicaud)), the function could be constrained to the background or the variables δu_β and δv_β could be left unchanged. The subscript β stands for β-plane hypothesis. In order to smooth the difference between the two cases, the velocity fields can be expressed as a weighted sum of the $(\delta u_\beta, \delta v_\beta)$ and $(\delta u_f, \delta v_f)$.

The dynamic height η is first described as a function of density and could also be more completely explained as a function of in-situ density and barotropic stream function ψ as proposed by Pinardi (1995) [5].

$$\delta\eta = \frac{f_o}{g}\frac{\delta\psi}{H} - \frac{1}{\rho_0 H}\int_{-H}^0 (z+H)\delta\rho dz \tag{66}$$

The establishment of the operator K is achieved through the following steps:

- the choice of a total variable,

- the formulation of the multivariate relationships between the balanced and unbalanced components of the state variables (f_{ij} functions),

- the computation of the coefficients in f_{ij} functions (statistics on the ensemble of runs).

5 The ensemble of forced-ocean simulations with OPA

5.1 Definition of an ensemble of forced simulations

As presented in the previous sections, the ensemble approach could be a way to estimate some elements of the background error covariance matrix $\mathbf{B}$ to which we can fit our covariance model.

The first idea is to work out several forced runs with OPA to calculate standard error deviation for each model variable of the state vector $\delta\mathbf{x}$ (the matrix $\mathbf{\Sigma_U}$ elements).

A reference run is carried out with an unperturb wind and SST forcing. Then additional simulations with perturbed wind forcing are performed. For the moment we use 4 perturbations on wind, the perturb runs are restarted from the reference run every 30 days.This time period is close to the data assimilation period for 4D VAR, it is chosen in order to have an idea of the grooth of the errors on the assimilation window. The SST forcing is left unchanged. These conditions can be adjusted to get a better dispersion of the ensemble.

For the univariate problem, one could have calculated the Σ elements straight from these results computing the standard error deviations for the model variables T, S, U, V and η depending on the free surface or rigid lid hypothesis.

In the multivariate case the operator $\mathbf{\Sigma}$, applies to the unbalanced components of the model variables only as explained in section 2: to access to the state vector $\delta\mathbf{x_U}$, the multivariate system $\mathbf{K}$ must be defined.

5.2 The correlations in space for the same model variable : the operator C

The constitution of the correlation operator $\mathbf{C}$ is based on two main ideas:

- First, the 3D problem is separated in a horizontal 2D diffusion case and a 1D vertical one.

- Then, the horizontal diffusion case solution is approximated with a Gaussian correlation function,which length scale is computed with a statistical approach.

5.3 Covariances between different model variables : the multivariate operator K and it's adjoint

As exposed in section (4), the multivariate aspect of $\mathbf{B}$ is represented by the operator $\mathbf{K}$ and it's adjoint. This operator determines the covariances between different variables of the model. These covariances lead to an impact on the analysis of all the model variables, even those not directly assimilated.

The forced-ocean simulations ensemble provides a large enough sample of to define the linear relationships between the variables.

Indeed, we have at our disposal several fields for each state vector variable and then we can find more or less complex physical and statistical relations between the variables.

For example, the relation (48) expresses a statistical law between the variables S and T, the balanced part of S is chosen as the part of the variable that strictly follows this statistical law, and the unbalanced part is the statistical error or residual.

6 Conclusion

A description of the background covariance matrix as a sequence of operators has been presented. The theoretical hypothesis of the description is based on Derber and Bouttier's paper ([3]) and lies on the separation of the background error covariance matrix in several operators. This partition implies the description of the increment vector as balanced and unbalanced components.

The univariate and multivariate aspects of the background error covariance matrix are separated, the univariate operators being applied to the unbalanced part of the increment vector. The univariate correlation operator is modeled considering the solution of a generalized diffusion equation on the sphere. The complete description of the method and algorithm is presented by A. Weaver and P. Courtier ([8]).

The multivariate notion appears in the formulation of the balance operator that links the balanced components of each variable of the increment vector to the unbalanced components of the other variables. Thus, the $T-S$ preservation hypothesis described in Troccoli and Haines's article ([2]) is directly included in the background error covariance matrix using a local an accurate linearization of $T - S$ relation from the background.

It should be emphasized that a good expression of the $T - S$ relation could, at first, improve results in a temperature assimilation context.

References

[1] A.Troccoli and Al. Salinity adjustements in the presence of temperature data assimilation. *Technical Memorandum 305, ECMWF*, 2000.

[2] A.Troccoli and K.Haines. Use of temperature-salinity relation in a data assimilation context. *American Meteorological Society*, 1999.

[3] J. Derber and F.Bouttier. A reformulation of the background error covariance in the ecmwf global data assimilation system. *Tellus*, 51A(195-221), 1999.

[4] F.Vossepoel. Use of sea level observations to estimate salinity variability in the tropical pacific. *J.Atmos.Oceanic Technol.*, 1999.

[5] N.Pinardi and A. Rosati. The sea surface pressure formulation of rigid lid model. implications for altimetric data assimilation studies. *Journal of Marine Systems*, 6(109-119), 1995.

[6] D. Parrish and J.Derber. The national meteorologocal center's spectral statistical interpolation analysis system. *Mondial Weather Revue*, 1992.

[7] P.Gauthier and M.Buehner. Background error statics modelling in a 3d variational data assimilation scheme:estimation and impact on the analyses. *Joint Statistics meetings*, 1998.

[8] A. Weaver and P.Courtier. Correlation modelling on the sphere using a generalized diffusion equation. *Q.J.Roy.Meteor.Soc*, 2000.