Optimal Spectral Decomposition (OSD) for Ocean Data Assimilation

Peter C. Chu*, Robin T. Tokmakian, and Chenwu Fan

Naval Ocean Analysis and Prediction Laboratory, Department of Oceanography
Naval Postgraduate School, Monterey, California, USA

Charles L. Sun
National Oceanographic Data Center, Silver Spring, Maryland, USA

*Corresponding author address: Department of Oceanography, Naval Postgraduate School, Monterey, CA 93943. Email: pcchu@nps.edu
Abstract

Optimal spectral decomposition (OSD) is applied to ocean data assimilation with variable (temperature, salinity, or velocity) anomalies (relative to background or modeled values) decomposed into generalized Fourier series, such that any anomaly is represented by a linear combination of products of basis-functions and corresponding spectral coefficients. It has three steps: (1) determination of the basis-functions, (2) optimal mode truncation, and (3) update of the spectral coefficients from innovation (or called observational increment). The basis-functions, depending only on the topography of the ocean basin, are the eigenvectors of the Laplace operator with the same lateral boundary conditions as the assimilated variable anomalies. The Vapnik-Chervonkis dimension is used to determine the optimal mode truncation. After that, the model field updates due to innovation through solving a set of linear algebraic equations of the spectral coefficients. The strength and weakness of the OSD method are demonstrated through a twin-experiment using the Parallel Ocean Program (POP) model.
1. Introduction

Data assimilation is required for operational ocean studies and maneuvers (Sun 1999), and has contributed significantly to the success of ocean modeling and prediction. In a numerical ocean model, a single variable or all the model variables \( c \) (no matter two or three dimensional) can be ordered by grid point and by variable, forming a single vector of length \( NP \) with \( N \) the total number of grid points and \( P \) the number of variables. For multiple model variables, non-dimensionalization is conducted before forming a single vector \( c \). The existing data assimilation is to blend modeled (or background) fields \( (c_b) \) (usually on the grid points) with observational data \( (c_o) \) (usually not at the grid points) of any ocean variable (Cohn, 1997; Tang and Kleeman 2004; Chu et al., 2004a; Galanis et al. 2006; Lozano et al., 2006),

\[
e_a = c_b + W[c_o - H(c_b)],
\]

(1)
to represent the (unknown) “truth” \( c_t \) with an analysis error,

\[
\varepsilon_a = c_a - c_t.
\]

(2)
Here, \( c_a \) is the assimilated field (or called analysis field); \( H \) is an operator that provides the model’s estimate at the observational points; \( W \) is the weight matrix; and \( d = [c_o - H(c_b)] \), is the innovation (or called the observational increment) (Fig. 1). Various data assimilation schemes such as optimal interpolation (OI), Kalman filter, and variational method (3DVAR and 4DVAR) were developed, and given unified notation by Ide et al. (1997). Their differences are the different ways to determine the weight matrix \( W \). For example, minimization of the cost function in the OI gives the weight matrix (e.g., Bretherton et al., 1976; Lozano et al., 2007),

\[
W = BH^T (HBH^T + R)^{-1}.
\]

(3)
The minimization of the analysis error covariance (P) in the Kalman filter (Galanis et al., 2006) leads to

\[ W_i = P^f(t_i)H_i^T[H_iP^f(t_i)H_i^T + R_i]^{-1}. \]  

(4)

Here, \( B \) and \( P^f \) are the background error covariance matrices; \( P^f \) is also called the forecast projection matrix by some authors; \( R_i \) is the observational error covariance matrix; and \( t \) is time. Despite some differences in formality, Eqs.(3)-(4) are identical. The most significant challenge for the existing data assimilation methods is the determination of the background error covariance matrix \( B \) (or forecast projection matrix \( P^f \)) for the OI and 3DVAR (or Kalman filter) since \( B \) and \( P^f \) are enormous matrices that are difficult to estimate due to the following characteristics: uncertain tunable parameters, inhomogeneous and anisotropic structures, complex boundaries in oceans.

In standard OI the covariance of a field between two events (space, time) or between a field at a grid location and an observation are prescribed from some general considerations on the nature of the covariances. These covariances can be converted to their equivalent representations in spectral space. Oceanographers have constructed \( B \) to include inhomogeneities and anisotropies associated to the presence of topography, and to reflect in a way the adaptation of the ocean fields to the topography. Utilization of ocean topography may change the weighting operation, \( W[c_o - H(c_b)] \) in (1), into a mathematical operator, \( F[c_o - H(c_b)] \), that maps the innovation (at the observational points) directly onto the grid points,

\[ c_a - c_b = F[c_o - H(c_b)] = F[\Delta c_o], \quad \Delta c_o = c_o - H(c_b), \]  

(5)

where \( H \) could be different in (1) and (5) when vertical interpolation is involved. The difference, \( \Delta c = c - c_b \), is called the anomaly (relative to the background field) of \( c \).
Very early in the application of OI to ocean fields Bretherton et al. (1976) explored the use of spectral representation of functions defined on a grid, instead of field values defined on a grid was considered. Along their path, the optimal spectral decomposition (OSD) was developed to apply the spectral method to field values, i.e., to perform as such an operator, $\mathcal{F}$, with the eigenvectors of the Laplace operator as the basis functions that only depend on the topography, satisfy the same boundary conditions as the assimilated ocean variable anomalies (e.g., temperature, salinity, velocity), and are pre-determined before the data assimilation.

Although the relative simplicity of an atmospheric spherical shell in comparison to the complexity of oceanic basins might explain the limited use of spectral models for the ocean, the OSD has been proven an effective ocean data analysis method. With it, several new ocean phenomena have been identified from observational data such as a bi-modal structure of chlorophyll-a with winter/spring (February–March) and fall (September–October) blooms in the Black Sea (Chu et al. 2005a), fall–winter recurrence of current reversal from westward to eastward on the Texas–Louisiana continental shelf from the current-meter, near-surface drifting buoy (Chu et. al 2005b), propagation of long Rossby waves at mid-depths (around 1000 m) in the tropical North Atlantic from the Argo float data (Chu et al. 2007), and temporal and spatial variability of global upper ocean heat content (Chu 2011) from the data of the Global Temperature and Salinity Profile Program (GTSPP, Sun et al. 2009). However, the OSD method has not yet been used for ocean data assimilation.

The purpose of this paper is to extend the use of OSD from ocean data analysis to ocean data assimilation. The OSD can be either three or two dimensional. However, it is
conducted in horizontal plane (i.e., two dimensional OSD) in this study. Rest of the paper is organized as follows. Section 2 discusses the lateral boundary conditions. Sections 3 describes the generation of basis-functions. Section 4 presents variables at grid versus observational points. Section 5 shows the determination of spectral coefficients from minimization of combined observational and analysis errors. Section 6 illustrates the mode truncation as a statistical learning process using the Vapnik-Chervonenkis (VC) dimension. Section 7 shows the ocean model with the OSD data assimilation procedure. Section 8 and 9 describe a twin experiment and error statistics of the OSD data assimilation. Section 10 presents the conclusions.

2. Lateral Boundary Condition

Let \((x, z)\) be the horizontal and vertical coordinates; \(R(z)\) be the area bounded by the lateral boundary \(\Gamma(z)\). The anomaly \(\Delta c\) satisfies the generalized homogeneous lateral boundary (\(\Gamma\)) condition (see Appendix for detail explanation),

\[
b_1(\tau) \mathbf{n} \nabla (\Delta c) + b_2(\tau) \Delta c = 0,
\]

where \(\nabla = \mathbf{i} \partial / \partial x + \mathbf{j} \partial / \partial y\) is the horizontal gradient operator with \((\mathbf{i}, \mathbf{j})\) the unit vectors in the horizontal plane; \(\mathbf{n}\) is the unit vector normal to the boundary; \(\tau\) denotes a moving point along the boundary, and \([b_1(\tau), b_2(\tau)]\) are parameters varying with \(\tau\). The boundary condition (6) becomes the Dirichlet boundary condition when \(b_1 = 0\), and the Neumann boundary conditions when \(b_2 = 0\). It is noted that different variable anomalies have different \([b_1(\tau), b_2(\tau)]\). For example, the temperature, salinity, and velocity potential anomalies have \(b_2 = 0\) for the rigid boundary and \(b_1 = 0\) for the open boundary. However, the streamfunction anomaly has \(b_1 = 0\) for the rigid boundary and \(b_2 = 0\) for the open boundary.
3. Basis Functions

3.1. Three Necessary Conditions

Selection of basis-functions \( \{ \phi_k(x, z) \} \) needs to satisfy three necessary conditions: (i) **satisfaction of the same homogeneous boundary condition** \( (6) \) of the assimilated variable anomaly, (ii) orthonormal, and (iii) independence on the assimilated variable. The second necessary condition is given by

\[
\iint_{R(z)} \phi_k(x, z)\phi_{k'}(x, z)dx = \delta_{kk'},
\]

where \( \delta_{kk'} \) is the Kronecker delta,

\[
\delta_{kk'} = \begin{cases} 
0 & \text{if } k \neq k' \\
1 & \text{if } k = k'. 
\end{cases}
\]

Due to independence on the assimilated variable (the third necessary condition), the basis-functions are available prior to the data assimilation.

Use of the eigenvectors of the horizontal Laplacian operator as the basis functions is an effective and easy way to get the basis-functions which satisfy the three necessary conditions. The eigenvectors \( \{ \phi_k \} \) of the horizontal Laplacian operator are the solutions of the Poisson equation,

\[
\nabla^2 \phi_k = -\lambda_k \phi_k, \quad [b_1(\tau)\mathbf{n} \cdot \nabla \phi_k + b_2(\tau)\phi_k]_{\Gamma} = 0, \quad k = 1, \ldots, \infty.
\]

Here, \( \{ \lambda_k \} \) are the eigenvalues, and \( \mathbf{n} \) is the unit vector normal to the lateral boundary. It is noted that these eigenvectors \( \{ \phi_k \} \) satisfy the three necessary conditions such as (i) satisfaction of the same homogeneous boundary condition \( (9) \) as the assimilated variable anomaly, (ii) orthonormal, and (iii) independent of the assimilated variables. The features (i) and (iii) distinguish the eigenvectors \( \{ \phi_k \} \) from the commonly used empirical
orthogonal functions (EOFs) in ocean data assimilation (e.g., Pham et al., 1998). The EOFs depend on the assimilated variables and do not satisfy the same homogeneous boundary condition (9) as the assimilated variable anomalies.

Due to irregular lateral boundaries, the basis-functions \( \{ \varphi_k \} \) are usually numerical solutions of (9), \( \{ \varphi_k(x_n, z) \} \). Here, \( x_n = (x_i, y_j) \), \( n = 1, 2, \ldots, N \), representing the horizontal grid points. From now on, the vertical coordinate \( z \) is omitted for simplicity. The first \( K \) discrete basis functions for all grid points are represented by the following matrix,

\[
G = \{ g_{nk} \} = \begin{bmatrix}
\varphi_1(x_1) & \varphi_1(x_2) & \ldots & \varphi_1(x_N) \\
\varphi_2(x_1) & \varphi_2(x_2) & \ldots & \varphi_2(x_N) \\
\ldots & \ldots & \ldots & \ldots \\
\varphi_K(x_1) & \varphi_K(x_2) & \ldots & \varphi_K(x_N)
\end{bmatrix}.
\] (10)

3.2. Example

With the NOAA National Geophysical Data Center’s Digital Bathymetry Data Base with 5’×5’ resolution (ETOPO5), the basis-functions \( \varphi_k(x_n) (k = 1, \ldots, K) \) at a certain depth \( z \) are computed for the Pacific Ocean. In assimilating temperature observations, the temperature anomaly \( \Delta c \) satisfies the Dirichlet boundary condition \( (b_1 = 0) \) at the southern boundary (Antarctic), and the Neumann boundary condition \( (b_2 = 0) \) elsewhere (rigid boundary). Fig. 2 shows the first 12 basis functions \( \{ \varphi_k \} \) for the Pacific Ocean at the surface for illustration. The first basis-function \( \varphi_1(x_n) \) shows the latitudinal variability. The second-basis function \( \varphi_2(x_n) \) shows the dipole-pattern of zonal variability with opposite signs in the eastern Pacific (negative) and the western Pacific (positive). The third basis-function \( \varphi_3(x_n) \) shows the slanted dipole-pattern with opposite signs in the
northeastern Pacific (positive) and the southwestern Pacific (negative). The fourth basis-
function $\phi_4(x_n)$ shows the tripole-pattern with negative values in the western and eastern
Pacific and positive values in between. The higher order basis functions have more
complicated variability structures. Some features are quite similar to recently described
global thermal structure (e.g., Chu, 2011). It may imply the topographic effect (at least
partially) on the horizontal variability such as temperature, salinity, density, and velocity
potential (Song et al. 2001).

4. Grid Versus Observational Points

Let variable $c$ have $M$ observations $c_o(x^{(m)}, t)$ at location $x^{(m)}$ ($m = 1, 2, \ldots, M$)
indicated by a superscript $(m)$, and have background field $c_b(x_n, t)$ at grid points $x_n$. The
$H$ operator in (1) is to interpolate the background (modeled) field ($c_b$) from grid to
observational point $x^{(m)}$,

$$
c_b(x^{(m)}, t) = \sum_{n=1}^{N} h_{mn} c_b(x_n, t),
$$

(11)

with

$$
\sum_{n=1}^{N} h_{mn} = 1.
$$

(12)

The innovation $d^{(m)}$ is then given by

$$
d^{(m)}(x^{(m)}, t) = c_o(x^{(m)}, t) - \sum_{n=1}^{N} h_{mn} c_b(x_n, t).
$$

(13)

Let $H = [h_{mn}]$ be an $M \times N$ matrix. Distribution of all innovations $d^{(m)}(x^{(m)})$ ($m = 1, 2, \ldots, M$) from the observational points into the grid points is represented by the same
proportionality coefficient. The mean adjustment at the grid point $x_n$ due to all the
observations is given by
where $f_n$ denotes observational data influence at the grid point $x_n$. The larger the value of $f_n$, the larger the observational influence on that grid point. Let $d$ be the $M$-dimensional innovation vector and $D$ be its distributed $N$-dimensional vector on the grid points,

\[ d^T = (d^{(1)}, d^{(2)}, \ldots, d^{(M)}), \quad D^T = (D_1, D_2, \ldots, D_N), \]

where the superscript $T$ indicates the transpose. Eq(14) can be written in matrix form,

\[ FD = H^T d, \]  \hspace{1cm} (16)

and $F$ is an $N \times N$ diagonal matrix

\[ F = \begin{bmatrix} f_1 & 0 & 0 & 0 \\ 0 & f_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & f_N \end{bmatrix}, \]  \hspace{1cm} (17)

which is called the data influence matrix. It is noted that both matrices $F$ and $H$ depend solely on the location of the observational points ($x^{(m)}$).

5. **OSD Data Assimilation Equation**

The observational error ($\epsilon_o$) at the grid point $x_n$ is given by

\[ \epsilon_o(x_n, t) = c_o(x_n, t) - c_i(x_n, t). \]  \hspace{1cm} (18)

The components of the vector $D$ represent the difference (observation minus background values) at the grid points,

\[ f_n D_n = c_o(x_n, t) - c_i(x_n, t). \]  \hspace{1cm} (19)

Its spectral form, $D_n$ is represented by
\[ D_n^{(K)} = \sum_{k=1}^{K} a_k(t) \phi_k(x_n), \quad (20) \]

where \( K \) is the mode number of the optimal truncation (see Section 6). The assimilated field with the given \( K \) is represented by

\[ c_a^{(K)}(x_n, t) = c_b(x_n, t) + f_n D_n^{(K)}, \quad (21) \]

The difference between \( D_n \) and \( D_n^{(K)} \) is given by

\[ f_n(D_n - D_n^{(K)}) = c_a(x_n, t) - c_a^{(K)}(x_n, t). \quad (22) \]

Substitution of (2) and (18) into (22) leads to

\[ f_n(D_n - D_n^{(K)}) = [c_a(x_n, t) - c_l(x_n, t)] - [c_a^{(K)}(x_n, t) - c_l(x_n, t)] = \varepsilon_o - \varepsilon_a^{(K)} \equiv \varepsilon^{(K)}, \quad (23a) \]

which contains the observational error (\( \varepsilon_o \)) and analysis error for a given \( K \) \([\varepsilon_a^{(K)}]\) at the grid point \( x_n \). It is noted that \( \varepsilon_o \) contains the instrumentation error especially those associated with remote sensing observations, and the representativeness error that is, for example, the mismatch of the point observation with the ocean model resolution. The combined observation-analysis error variance over the whole domain is defined by

\[ J_{tr} = \frac{\sum_{n=1}^{N} [f_n(D_n - D_n^{(K)})]^2}{N - 1} \quad (23b) \]

Minimization of \( J_{tr} \) after substituting (19) and (20) into (23b),

\[ \frac{\partial J_{tr}}{\partial a_k} = \frac{1}{N - 1} \frac{\partial}{\partial a_k} \left\{ \sum_{n=1}^{N} \sum_{k=1}^{K} [f_n(t)\phi_k(x_n) - D_n(t)]^2 \right\} = 0 \]

leads to a set of algebraic equations for the spectral coefficients \( \{a_k\} \),

\[ \sum_{k=1}^{K} \left( \sum_{n=1}^{N} \phi_l(x_n) f_n \phi_k(x_n) \right) a_k = \sum_{n=1}^{N} \phi_l(x_n) f_n D_n, \quad l = 1, 2, \ldots, K, \quad (24) \]

which is rewritten into a matrix form after using (10),
\textbf{GFG}^T \mathbf{A} = \mathbf{GFD}, \quad (25)

where \( \mathbf{A} \) is the spectral coefficient vector, \( \mathbf{A}^T = (a_1, a_2, \ldots, a_K) \). The solution of (25) is given by

\[
\mathbf{A} = \left[ \text{GFG}^T \right]^{-1} \mathbf{GFD}. 
\quad (26)
\]

Eq. (21) is then written into the matrix form after using (16),

\[
\mathbf{c}_a = \mathbf{c}_b + \text{FG}^T \left[ \text{GFG}^T \right]^{-1} \mathbf{E}, \quad \mathbf{E} \equiv \mathbf{GFD} = \mathbf{GH}^T \mathbf{d}, \quad (27)
\]

which is called the OSD data assimilation equation. The vector \( \mathbf{E} \) denotes the observational innovation projected into the spectral space. The matrix form of the OSD data assimilation is quite similar to the existing ocean data assimilation schemes (OI and Kalman filter with the matrix \( \mathbf{B} \) replaced by \( \mathbf{F} \) and \( \mathbf{H} \) by \( \mathbf{G} \) [see (3) and (4)]. The two matrices \( \mathbf{B} \) and \( \mathbf{F} \) play a similar role that make the analysis field compact in the observational data-rich area. It is also noted that the OSD data assimilation equation (27) is applied to one vertical level \( z \). As such, the data assimilation may distort the vertical stratification. Recently developed fully conserved minimal adjust schemes (Chu and Fan, 2010; Wang et al., 2012) can be used to stabilize the vertical stratification.

6. Mode Truncation Using the Vapnik-Chervonenkis Dimension

The assimilation results depend on the mode truncation \( K \) since the spectral coefficients \( (a_1, \ldots, a_K) \) are determined on the base of minimization of the combined observation-analysis error variance \( J_{tr} \) [see (23b)] for the given \( K \). With the calculated spectral coefficients \( (a_1, \ldots, a_K) \) based on observational data, the assimilated field, \( \mathbf{c}_a^{(K)}(\mathbf{x}_n,t) \), can be calculated at any grid point \( (\mathbf{x}_n) \) using (27), and the analysis error is estimated by [see (2)]
\( \varepsilon_a^{(K)}(\mathbf{x}_n,t) = \varepsilon_b(\mathbf{x}_n,t) + \sum_{k=1}^{K} a_k(t) \phi_k(\mathbf{x}_n), \quad \varepsilon_b(\mathbf{x}_n,t) \equiv c_b(\mathbf{x}_n,t) - c_t(\mathbf{x}_n,t), \) \hspace{1cm} (28)

where \( \varepsilon_b(\mathbf{x}_n,t) \) is the model (or background) error. It is noted from (28) that reduction of the model (or background) error \( \varepsilon_b \) (i.e., smaller \( \varepsilon_a^{(K)} \)) is achieved by the observational innovation using OSD (second term in the right-hand side).

Since the “true” field, \( c_t(\mathbf{x}_n,t) \), is still uncertain, the analysis error \( \varepsilon_a^{(K)} \) should be estimated probabilistically. In the spectral decomposition method the observation space and the model space are projected into the spectral space. There is a need to ensure that the size of the spectral space is adequate for these two purposes. The spectral representation acts as a spatial low pass filter for the fields, where the highest allowed wavenumbers relate to the highest spectral eigenvalues. The observational network is required to provide information without aliasing. For example, in an eddy field in the deep ocean one expects that the basis functions can resolve well features of the size of the Rossby radius of deformation. Thus, the ratio of observational points \((M)\) and the spectral truncation \((K)\) is a key to determine the optimal mode truncation \(K_{opt}\). It is noted that \( c_b(\mathbf{x}_n,t), (a_1, \ldots, a_K), \) and \( \{ \phi_k, k =1, \ldots, K \} \) are given. Let \( J \) be the ensemble average of analysis error variance. The probability for the upper bound of \( J \) is given by (Vapnik 2000; Chu et al., 2003a)

\[
P \left\{ J \leq J_{tr} + J_* \sqrt{\frac{\ln(2M / K) + 1}{M / K}} - \ln(\eta / 4) \right\} = 1 - \eta,
\]

(29)

where the mode truncation \( K \) is treated as the Vapnik-Chervonenkis (VC) dimension; and \( \eta (\ll 1) \) is the significance level. \( J_* \) is the upper bound of \( J_{tr} \). The minimization of the VC cost function \( (J_K) \),
\[ J_K = J_{tr} + \mu(K, M, \eta), \]
\[ \mu(K, M, \eta) = J_{tr} \left( \sqrt{\frac{\ln\left(\frac{2M}{K} + 1\right) - \ln(\eta / M)}{M / K}} \right), \quad K = 1, 2, ..., \infty \]  
(30)

leads to another set of spectral coefficients \(\{a_1^*, ..., a_K^*\}\). It is noted that for a given \(M\), \(J_{tr}\) decreases monotonically with \(K\); \(\mu\) increases with \(K\) if \(\eta\) is given (\(\eta = 0.1\) in this study). Thus, \(J_K\) has a minimum value for certain mode number \(K_{opt}\),

\[ \min_K (J_K) = J_{K_{opt}}. \]  
(31)

7. Ocean Modeling

7.1. Model Description

The Parallel Ocean Program (POP) model is used to show the feasibility of the OSD data assimilation. Within the framework of Community Earth System Model (CESM), the POP is a time-dependent, level-coordinate, primitive equation ocean general circulation model rendered on a three-dimensional grid that includes a free surface and realistic topography (Smith et al. 2010). The B-grid is used for the spatial discretization. Derived from the Bryan-Cox-Semtner class of models, the POP (Dukowicz and Smith 1994) was officially adopted as the ocean component of the CESM based at NCAR in 2001. It has an implicit free surface and general orthogonal coordinates (Smith et al., 1995). It is a global model, with the grid defined so that the pole is located in Greenland. Since the purpose of this study is to show the feasibility of the OSD data assimilation rather than to simulate/predict the real ocean processes, a low horizontal resolution (3°), 60 vertical levels (Table 1) version of the model with a time step of 2 hours is used in this study. In the top 175 m, the model has 30 levels with 10 m between each of the consecutive levels. The discretized model variable at the grid points is represented by
\( c(x_n, z_l, t), n = 1, 2, \ldots, N_l, l = 1, 2, \ldots, L. \) Here, \( N_l \) is the total number of the horizontal grid points at the vertical level \( l; L = 60, \) is the total number of the vertical levels.

The atmospheric forcing at the surface is provided by an annually varying climatology derived from the surface Coordinated Ocean Research Experiments (CORE) version 2 (Large and Yeager 2009). The air–sea fluxes of momentum, heat, freshwater and their components have been computed globally from 1948 at frequencies ranging from 6-hourly to monthly. All fluxes are computed over the 23 years from 1984 to 2006, but radiation prior to 1984 and precipitation before 1979 are given only as climatological mean annual cycles. The input data are based on NCEP reanalysis for the surface vector wind, temperature, specific humidity and density, and on a variety of satellite based radiation, sea surface temperature, sea-ice concentration and precipitation products (from the website: https://climatedataguide.ucar.edu/climate-data/large-yeager-air-sea-surface-flux-corev2-1949-2006). The model simulations for this experiment used climatological forcing, (daily 23 year average). The forcing is interpolated to the time step of the model.

The POP model has been spun up from rest and climatological annual mean \((T, S)\) with the daily climatological surface forcing from the CORE version 2 (Large and Yeager 2009) and integrated for a period of over 300 simulation years. The model output for the year-300 \((c_{300})\) is treated as the “truth field”, \( c_i(x_n, z_l, t) \).

### 7.2. Initial Error

Although we are using a global model, temperature “observations” are only incorporated for the Pacific basin. It is noted that use of single variable (i.e., temperature) data is not ideal since observational temperature \((T)\), salinity \((S)\), and velocity \((V)\) data should be assimilated to keep dynamic balance since \((T, S, V)\) are the dependent variables
in ocean models. Chu (2006) shows that assimilation with \((T, S)\) data only introduces
dynamic imbalance and suggests geostrophic velocity corresponding to the \((T, S)\) data
should also be assimilated. The results of this study are only used for the preliminary
evaluation.

The model is integrated from 1 March of year-210 and using “observations”
sampled from the fields from 1 March of year-300. The initial error (the variable \(c\)
denoting temperature) is

\[
e(x_n, z_l, t_0) = c_{210}(x_n, z_l, t_0) - c_{300}(x_n, z_l, t_0).
\] (32)

The temperature at the surface initially has maximum errors (i.e., the mismatch between
years 210 and 300) such as \(+2^\circ C\) in the Southern Ocean near Antarctic, \(-2^\circ C\) north of the
Kuroshio extension, medium errors such as \(+1^\circ C\) in the central equatorial Pacific, and
low errors (\(|e| < 0.5^\circ C\)) in sub-tropical areas in both hemispheres (Fig. 3a). The
temperature initially has smaller errors at 1106 m depth (level-41) with maximum errors
in the circumpolar currents with near \(+1^\circ C\) in the west and \(-1^\circ C\) in the east, and low
errors (\(|e| < 0.5^\circ C\)) elsewhere (Fig. 3b). The model without data assimilation is
integrated from 1 March with the initial condition,

\[
c_b(x_n, z_l, t_0) = c_{210}(x_n, z_l, t_0),
\] (33)

using daily surface forcing for 20 days, represented by \(c_{non}(x_n, z_l, t)\).

8. OSD Data Assimilation

8.1. Bilinear Interpolation
Let the observational point \( x^{(m)} \) located in the grid cell \( x_i \leq x^{(m)} < x_{i+1} \), \( y_j \leq y^{(m)} < y_{j+1} \). In this study, the bilinear interpolation is chosen for the H-operator (Fig. 4),

\[
H(c_k) = w_{i,j}^{(m)} c_{b(i,j)} + w_{i+1,j}^{(m)} c_{b(i+1,j)} + w_{i,j+1}^{(m)} c_{b(i,j+1)} + w_{i+1,j+1}^{(m)} c_{b(i+1,j+1)},
\]

where

\[
w_{i,j}^{(m)} = \frac{(x_{i+1} - x^{(m)})(y_{j+1} - y^{(m)})}{(x_{i+1} - x_i)(y_{j+1} - y_j)}, \quad w_{i+1,j}^{(m)} = \frac{(x^{(m)} - x_i)(y_{j+1} - y^{(m)})}{(x_{i+1} - x_i)(y_{j+1} - y_j)}
\]

\[
w_{i,j+1}^{(m)} = \frac{(x_{i+1} - x^{(m)})(y^{(m)} - y_j)}{(x_{i+1} - x_i)(y_{j+1} - y_j)}, \quad w_{i+1,j+1}^{(m)} = \frac{(x^{(m)} - x_i)(y^{(m)} - y_j)}{(x_{i+1} - x_i)(y_{j+1} - y_j)}.
\]

It is noted that the proportionality coefficients \( \{w_{i,j}^{(m)}, w_{i+1,j}^{(m)}, w_{i,j+1}^{(m)}, w_{i+1,j+1}^{(m)}\} \) depend solely on the location of the observational points \( x^{(m)} \), and

\[
w_{i,j}^{(m)} + w_{i+1,j}^{(m)} + w_{i,j+1}^{(m)} + w_{i+1,j+1}^{(m)} = 1.
\]

Each row of the \( M \times N \) matrix \( H = [h_{mn}] \) in (16) only has 4 non-zero values,

\[
H = \begin{bmatrix}
0 & 0 & w_{11}^{(1)} & w_{21}^{(1)} & 0 & \ldots & w_{12}^{(1)} & w_{22}^{(1)} & 0 & \ldots & 0 & 0 \\
0 & w_{11}^{(2)} & w_{21}^{(2)} & \ldots & w_{12}^{(2)} & w_{22}^{(2)} & 0 & 0 & 0 & 0 & \ldots & 0 \\
0 & 0 & 0 & w_{11}^{(3)} & w_{21}^{(3)} & 0 & \ldots & w_{12}^{(3)} & w_{22}^{(3)} & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & w_{11}^{(M-1)} & w_{21}^{(M-1)} & \ldots & 0 & w_{12}^{(M-1)} & w_{22}^{(M-1)} & 0 & 0 & 0 & \ldots & 0 \\
0 & 0 & 0 & w_{11}^{(M)} & w_{21}^{(M)} & 0 & \ldots & w_{12}^{(M)} & w_{22}^{(M)} & 0 & 0 & 0
\end{bmatrix}.
\]

Other simple interpolations such as inverse distance weighting, spline, trigonometric polynomials can also be used for \( H \) matrix.
8.2. Twin Experiment

Sampling pattern consisting of data-rich area and data-poor area offers a challenge to the ability of the basis function to represent the intended fields well since the projection of the data onto the spectral fields is likely to generate a field defined in the entire domain. To test the ability of the OSD to represent the intended fields, the “observational” data is sampled from $c_{300}$ beginning with 1 March for 20 days at locations (unchanged during the data assimilation process) given by the horizontal distribution of the Argo floats in March 2003 (Fig. 5). This produces the “observational” data set $[c_a(x^{(1)}_n, z, t), ..., c_a(x^{(M)}_n, z, t)]$ with data-rich area north of 20°S and data-poor area south of 20°S. If the spatial decorrelation scale is much less that the domain size, the analysis fields using OI will be compact in the data-rich area (i.e., north of 20°S).

The OSD data assimilation process at day $= t$ follows (27) with the following procedure: (a) determine the optimal mode decomposition ($K_{opt}$); (b) compute the difference between the “observational” and background values at the observational points following Eq.(13), (c) substitute the difference into Eq.(26) to obtain the spectral coefficients [$a_1(z, t), a_2(z, t), ..., a_{Kopt}(z, t)$]; (d) substitute the spectral coefficients [$a_1(z, t), a_2(z, t), ..., a_K(z, t_0)$] into (27) to get the assimilated initial condition $c_a(x_n, z, t)$. The dependence of the VC cost function ($J_K$) on the VC dimension (Fig. 6) shows that the optimal mode truncation is, $K_{opt} = 12$ at 125 m depth (level-13, Table 1) and day-0 (for illustration). The assimilation model is then run forward in time for 24 hours with the model field saved at the end of 24 hours, which is the background field for the day $= (t+1)$, $c_b(x_n, z, t+1)$ day). At each assimilation time, the optimal mode truncation ($K_{opt}$) is
re-calculated. This process repeats for 20 days, and leads to the assimilated output, $c_a(x_n, z, t)$.

9. Error Statistics

The three datasets: $c_a(x_n, z_l, t)$, $c_{non}(x_n, z_l, t)$, $c_l(x_n, z_l, t)$ ($l = 1, \ldots, L$) for the period of 20 days ($t_0 \leq t \leq t_1 = t_0 + 20$ days) are used to show the root mean square error (RMSE) with and without the OSD data assimilation. Let $N_l$ be the number of the horizontal grid points at the vertical level $l$ and $L$ be the number of the total vertical levels ($L = 60$). The basin-wide RMSE and BIAS for the assimilation run ($E_{ass}, B_{ass}$) and non-assimilation run ($E_{non}, B_{non}$) are given by

$$E_{ass}(t) = \sqrt{\frac{1}{L} \sum_{l=1}^{L} \frac{1}{N_l} \sum_{i=1}^{N_l} \left[ c_a(x_i, z_l, t) - c_i(x_i, z_l, t) \right]^2}, \quad (37a)$$

$$E_{non}(t) = \sqrt{\frac{1}{L} \sum_{l=1}^{L} \frac{1}{N_l} \sum_{i=1}^{N_l} \left[ c_{non}(x_i, z_l, t) - c_i(x_i, z_l, t) \right]^2}, \quad (37b)$$

$$B_{ass}(t) = \frac{1}{L} \sum_{l=1}^{L} \frac{1}{N_l} \sum_{i=1}^{N_l} \left[ c_a(x_i, z_l, t) - c_i(x_i, z_l, t) \right], \quad (38a)$$

$$B_{non}(t) = \frac{1}{L} \sum_{l=1}^{L} \frac{1}{N_l} \sum_{i=1}^{N_l} \left[ c_{non}(x_i, z_l, t) - c_i(x_i, z_l, t) \right]. \quad (38b)$$

Fig. 7 shows the comparison of the basin-wide RMSE and BIAS of the model between without data assimilation (dashed curve) and with the OSD data assimilation (solid curve). RMSE increases from 0.50°C at day-1 to 0.52°C at day-20 (0.02°C increase) without data assimilation (4% of error increase); and decreases from 0.50°C at day-1 to 0.43°C at day-20 with the OSD data assimilation (14% of error decrease). BIAS increases from 0.080°C at day-1 to 0.082°C at day-20 (0.002°C increase) without data
assimilation (2.5% increase); and decreases from 0.08°C at day-1 to 0.04°C at day-20 (0.04°C decrease) with the OSD data assimilation (50% decrease).

The local RMSE and BIAS for the assimilation and non-assimilation runs are given by

\[
E_{\text{ass}}(x_n,t) = \frac{1}{L} \sum_{j=1}^{L} [c_a(x_n,z_j,t) - c_j(x_n,z_j,t)]^2, \quad (39a)
\]

\[
E_{\text{non}}(x_n,t) = \frac{1}{L} \sum_{j=1}^{L} [c_{\text{non}}(x_n,z_j,t) - c_j(x_n,z_j,t)]^2, \quad (39b)
\]

\[
B_{\text{ass}}(x_n,t) = \frac{1}{L} \sum_{j=1}^{L} [c_a(x_n,z_j,t) - c_j(x_n,z_j,t)], \quad (40a)
\]

\[
B_{\text{non}}(x_n,t) = \frac{1}{L} \sum_{j=1}^{L} [c_{\text{non}}(x_n,z_j,t) - c_j(x_n,z_j,t)]. \quad (40b)
\]

Comparison of the local RMSE (Fig. 8) and BIAS (Fig. 9) between without data assimilation (right panels) and with the OSD data assimilation (left panels) for day-1 (upper panels), day-10 (middle panels), and day-20 (lower panels) shows the strength and weakness of the OSD scheme. At day-1, the local RMSE and BIAS are quite comparable between the assimilated run (Figs. 8a and 9a) and the non-assimilated run (Figs. 8b and 9b). The local RMSE has large values around (~2°C) in the central equatorial Pacific (10°S-10°N, 160°W-120°W), and the eastern tropical north Pacific (10°N-18°N, 120°W-90°W), and very narrow strip in the Antarctic Circumpolar Current region near the ice shelf (south of 68°S, 160°W – 90°W), and relatively low values elsewhere. The local BIAS has large values around (0.5~1°C) in the most areas of the low latitudes (20°S-20°N) and high latitudes (north of 40°N) except in the eastern Pacific near coast, and in the Antarctic Circumpolar Current region, and relatively low values elsewhere.
As the time progresses, the local (RMSE, BIAS) for the non-assimilate run remain almost the same at day-10 (Fig. 8d, Fig. 9d) and day-20 (Fig. 8f, Fig. 9f) as at day-1 (Fig. 8b, Fig. 9b), but changes evidently for the assimilated run at day-10 (Fig. 8c, Fig. 9c) and day-20 (Fig. 8e, Fig. 9e) as compared to day-1 (Fig. 8a, Fig. 9a).

The local RMSE is reduced drastically north of 20°S with disappearance of high local RMSE originally (day-1) in the central equatorial Pacific, and the eastern tropical north Pacific, and is not reduced or may even be increased slightly south of 20°S with the appearance of high local RMSE originally (day-1) in the Antarctic Circumpolar Current region near the ice shelf (south of 68°S, 160°W – 90°W). It is noted that the areas with large reduction in error and bias are the “observational” data-rich areas and with less decreasing (or even increasing) in error and bias are the “observational” data-poor areas (comparing Figs. 8 and 9 to Fig. 5).

10. Conclusions

The OSD method has been developed for ocean data assimilation on the base of the classic theory of the generalized Fourier series expansion such that any ocean field is represented by linear combination of the products of basis functions (or called modes) and corresponding spectral coefficients. The basis functions are the eigenvectors of the Laplace operator, determined only by the topography with the same lateral boundary conditions for the assimilated variables.

Different from the existing ocean data assimilation methods such as optimal interpolation, Kalman filters, and variational methods (originally developed for atmospheric data assimilation), the OSD method has the four specific features: (a) effective utilization of the ocean topographic data, (b) orthonormal and predetermined
basis functions which are independent on and satisfy the same lateral boundary condition of the assimilated variable anomalies, (c) no requirement of a priori information on a background error covariance matrix $B$, (d) optimal mode truncation through minimization of the Vapnik-Chervonkis dimension as a statistical learning process. After the mode truncation, the model field updates due to innovation through solving a set of a linear algebraic equations of the spectral coefficients.

The capability of the OSD method is demonstrated through a twin-experiment using the Parallel Ocean Program (POP) model for the Pacific Ocean. For an objective evaluation, the “observational” data are not uniformly distributed with data-rich area north of $20^\circ S$ and data-poor area south of $20^\circ S$. Within 20 days, the basin-wide RMSE (BISE) increases 4% (2.5%) without the OSD data assimilation, and decreases 14% (50%) with the OSD data assimilation. However, the improvement using the OSD data assimilation depends on the “observational” data distribution. The local RMSE is reduced drastically in data-rich areas (i.e., north of $20^\circ S$) but not in data-poor areas (i.e., south of $20^\circ S$).

No use of a-priori $B$ matrix implies that the observations are purely extrapolated to the data-poor area with the control of the observational influence matrix $F$ [see (17) and (27)]. Since the extrapolation causes unpredictable analysis errors and the twin-experiment does not show improvement by OSD assimilation in data-poor area, further studies on constructing the $F$ matrix are needed. Moreover, verification using a twin-experiment is just a first step. Feasibility studies should be conducted for real ocean data such as conductivity–temperature–depth (CTD), expendable bathythermograph (XBT), Argo profiling data, and glider data.
The OSD method proposed here is two-dimensional and conducted at each vertical level with the basis functions given by the eigenvectors of the horizontal Laplacian operator. This can be extended to a three-dimensional OSD method with the basis functions given by the eigenvectors of the three-dimensional Laplacian operator, where much larger matrix operations will be involved. Besides, for the three dimensional OSD, the surface boundary conditions of the assimilated variable anomalies may vary due to local climatology. Its impact on the three dimensional basis functions will be investigated in future studies.

**Acknowledgments.** The Office of Naval Research, the Naval Oceanographic Office, and the Naval Postgraduate School supported this study.

**Appendix. Derivation of Lateral Boundary Condition (6)**

Generally, the assimilated ocean variable $c$ (temperature, salinity, density, velocity, …) have the lateral boundary ($\Gamma$) condition,

$$b_1(\tau)n \cdot \nabla c + b_2(\tau)c = D(\tau),$$

(A1)

where $D(\tau)$ is the forcing term varying with $\tau$. With the inhomogeneous boundary condition (A1), the assimilated variable $c(x, z, t)$ consists of two parts,

$$c(x, z, t) = \hat{c}(x, z, t) + S(x, z, t),$$

(A2)

where $S(x, z, t)$ is the solution of the Laplace equation with the inhomogeneous boundary condition,

$$\nabla^2 S = 0, \quad k_1(\tau)n \cdot \nabla S + k_2(\tau)S = D(\tau) \text{ at } \Gamma;$$

(A3)

and $\hat{c}(x, z, t) = c(x, z, t) - S(x, z, t)$, satisfies the homogeneous boundary condition,
\[ b_1(\tau \mathbf{n} \cdot \nabla \hat{c} + b_2(\tau) \hat{c} = 0. \]  \hspace{1cm} (A4)

Since
\[ c_b(x, z, t) = \hat{c}_b(x, z, t) + S(x, z, t), \]  \hspace{1cm} (A5)

Subtraction of (A5) from (A2) leads to
\[ \Delta c(x, z, t) = c(x, z, t) - c_b(x, z, t) = \hat{c}(x, z, t) - \hat{c}_b(x, z, t). \]  \hspace{1cm} (A6)

Both \( \hat{c}(x, z, t) \) and \( \hat{c}_b(x, z, t) \) satisfy the boundary condition (A4), which leads to the boundary condition (6) for \( \Delta c \),

\[ b_1(\tau \mathbf{n} \cdot \nabla (\Delta c) + b_2(\tau) \Delta c = 0. \]

References


Table 1. Depths of vertical levels in the POP model.

<table>
<thead>
<tr>
<th>Level</th>
<th>Depth (m)</th>
<th>Level</th>
<th>Depth (m)</th>
<th>Level</th>
<th>Depth (m)</th>
<th>Level</th>
<th>Depth (m)</th>
<th>Level</th>
<th>Depth (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>13</td>
<td>125</td>
<td>25</td>
<td>268</td>
<td>37</td>
<td>708</td>
<td>49</td>
<td>2649</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>14</td>
<td>135</td>
<td>26</td>
<td>285</td>
<td>38</td>
<td>787</td>
<td>50</td>
<td>2890</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>15</td>
<td>145</td>
<td>27</td>
<td>305</td>
<td>39</td>
<td>879</td>
<td>51</td>
<td>3133</td>
</tr>
<tr>
<td>4</td>
<td>35</td>
<td>16</td>
<td>155</td>
<td>28</td>
<td>328</td>
<td>40</td>
<td>985</td>
<td>52</td>
<td>3380</td>
</tr>
<tr>
<td>5</td>
<td>45</td>
<td>17</td>
<td>165</td>
<td>29</td>
<td>351</td>
<td>41</td>
<td>1106</td>
<td>53</td>
<td>3628</td>
</tr>
<tr>
<td>6</td>
<td>55</td>
<td>18</td>
<td>175</td>
<td>30</td>
<td>378</td>
<td>42</td>
<td>1245</td>
<td>54</td>
<td>3876</td>
</tr>
<tr>
<td>7</td>
<td>65</td>
<td>19</td>
<td>186</td>
<td>31</td>
<td>409</td>
<td>43</td>
<td>1401</td>
<td>55</td>
<td>4126</td>
</tr>
<tr>
<td>8</td>
<td>75</td>
<td>20</td>
<td>198</td>
<td>32</td>
<td>443</td>
<td>44</td>
<td>1574</td>
<td>56</td>
<td>4375</td>
</tr>
<tr>
<td>9</td>
<td>85</td>
<td>21</td>
<td>210</td>
<td>33</td>
<td>483</td>
<td>45</td>
<td>1764</td>
<td>57</td>
<td>4625</td>
</tr>
<tr>
<td>10</td>
<td>95</td>
<td>22</td>
<td>223</td>
<td>34</td>
<td>528</td>
<td>46</td>
<td>1969</td>
<td>58</td>
<td>4875</td>
</tr>
<tr>
<td>11</td>
<td>105</td>
<td>23</td>
<td>236</td>
<td>35</td>
<td>579</td>
<td>47</td>
<td>2186</td>
<td>59</td>
<td>5125</td>
</tr>
<tr>
<td>12</td>
<td>115</td>
<td>24</td>
<td>251</td>
<td>36</td>
<td>639</td>
<td>48</td>
<td>2414</td>
<td>60</td>
<td>5375</td>
</tr>
</tbody>
</table>
Figure Captions

Fig. 1. Illustration of ocean data assimilation with $c_b$ located at the grid points, and $c_o$ located at the points “*”. The ocean data assimilation is to convert the innovation, $d = c_o - H(c_b)$, from the observational points to the grid points.

Fig. 2. First 12 second-type basis functions $\{\phi_k, k = 1,\ldots, 12\}$ for the Pacific Ocean at the surface.

Fig. 3. Initial errors in temperature (°C) at: (a) the sea surface, and (b) the depth of 1106 m (at the level 41).

Fig. 4. The bilinear interpolation for calculating the basis functions at the observational point $x_m$ from their values at the four neighboring grid points.

Fig. 5. Daily sampling taking from horizontal distribution of the Argo floats in March 2003. It is noted that the “observational” data-rich area is north of 20°S, and the “observational” data-poor area is south of 20°S.

Fig. 6. Optimal mode decomposition ($K_{opt}$) at 125 m depth and day-0 is determined the minimization of the VC cost function (denoted by red square).

Fig. 7. Comparison between the assimilation and non-assimilation runs of the temporally varying basin-wide (a) RMSE and (b) BIAS.

Fig. 8. Comparison of temporally varying local RMSE for the assimilation and non-assimilation runs (a) assimilation run at day-1, (b) non-assimilation run at day-1, (c) assimilation run at day-10, (d) non-assimilation run at day-10, (e) assimilation run at day-20, (f) non-assimilation run at day-20.

Fig. 9. Comparison of temporally varying local BIAS for the assimilation and non-assimilation runs (a) assimilation run at day-1, (b) non=assimilation run at day-1, (c) assimilation run at day-10, (d) non-assimilation run at day-10, (e) assimilation run at day-20, (f) non-assimilation run at day-20.
Fig. 1. Illustration of ocean data assimilation with $c_b$ located at the grid points, and $c_o$ located at the points "*". The ocean data assimilation is to convert the innovation, $d = c_o - H(c_b)$, from the observational points to the grid points.
Fig. 2. First 12 second-type basis functions \( \{ \phi_k, k = 1, \ldots, 12 \} \) for the Pacific Ocean at the surface.
Fig. 3. Initial errors in temperature (°C) at: (a) the sea surface, and (b) the depth of 1106 m (at the level 41).
Fig. 4. The bilinear interpolation for calculating the basis functions at the observational point $\mathbf{x}^{(m)}$ from their values at the four neighboring grid points.
Fig. 5. Daily sampling taking from horizontal distribution of the Argo floats in March 2003. It is noted that the “observational” data-rich area is north of 20°S, and the “observational” data-poor area is south of 20°S.
Fig. 6. Optimal mode decomposition ($K_{opt}$) at 125 m depth and day-0 is determined the minimization of the VC cost function (denoted by red square).
Fig. 7. Comparison between the assimilation and non-assimilation runs of the temporally varying basin-wide (a) RMSE and (b) BIAS.
Fig. 8. Comparison of temporally varying local RMSE for the assimilation and non-assimilation runs (a) assimilation run at day-1, (b) non-assimilation run at day-1, (c) assimilation run at day-10, (d) non-assimilation run at day-10, (e) assimilation run at day-20, (f) non-assimilation run at day-20.
Fig. 9. Comparison of temporally varying local BIAS for the assimilation and non-assimilation runs (a) assimilation run at day-1, (b) non-assimilation run at day-1, (c) assimilation run at day-10, (d) non-assimilation run at day-10, (e) assimilation run at day-20, (f) non-assimilation run at day-20.