

A Statistical Interpolation Code for Ocean Analysis and Forecasting

ASHWANTH SRINIVASAN,^a T. M. CHIN,^b E. P. CHASSIGNET,^c M. ISKANDARANI,^d AND N. GROVES^e

^a Tendral LLC, Miami, Florida

^b Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California

^c Center for Ocean Atmosphere Prediction Studies, Florida State University, Tallahassee, Florida

^d Rosenstiel School of Marine and Atmospheric Sciences, University of Miami, Miami, Florida

^e Woods Hole Group, Inc., Bourne, Massachusetts

(Manuscript received 22 March 2021, in final form 23 November 2021)

ABSTRACT: We present a data assimilation package for use with ocean circulation models in analysis, forecasting, and system evaluation applications. The basic functionality of the package is centered on a multivariate linear statistical estimation for a given predicted/background ocean state, observations, and error statistics. Novel features of the package include support for multiple covariance models, and the solution of the least squares normal equations either using the covariance matrix or its inverse—the information matrix. The main focus of this paper, however, is on the solution of the analysis equations using the information matrix, which offers several advantages for solving large problems efficiently. Details of the parameterization of the inverse covariance using Markov random fields are provided and its relationship to finite-difference discretizations of diffusion equations are pointed out. The package can assimilate a variety of observation types from both remote sensing and in situ platforms. The performance of the data assimilation methodology implemented in the package is demonstrated with a yearlong global ocean hindcast with a $1/4^\circ$ ocean model. The code is implemented in modern Fortran, supports distributed memory, shared memory, multicore architectures, and uses climate and forecasts compliant Network Common Data Form for input/output. The package is freely available with an open source license from www.tendral.com/tsis/.

KEYWORDS: Ocean circulation; Hindcasts; Operational forecasting; Data assimilation; Numerical weather prediction/forecasting

1. Introduction

With the rapid development and deployment of observing systems unprecedented amount of information is becoming available for oceanic process studies, ocean-state forecasting, and climate monitoring. The new data encompass measurements that sample the ocean at different spatial and temporal scales and are complemented by a fast-growing capacity of numerical models to simulate ocean processes across scales (Morrow et al. 2019; Fox-Kemper et al. 2019). To make use of these sources of information, it is often useful to combine them and derive estimates of the ocean state that are regular in spatial and temporal coverage, and simultaneously consistent with both measurements and models. Data assimilation, essentially methodologies to optimally merge information from different sources, has emerged as a powerful tool to integrate vast amounts of data in recent decades (Moore et al. 2019; Heimbach et al. 2019). Although data assimilation algorithms are relatively well established and many datasets are routinely available, using them for a particular task can prove to be a rather complicated undertaking. General purpose tools such as the Data Assimilation Research Testbed (DART; Anderson et al. 2009), the Data Interpolating Variational Analysis (DIVA; Troupin et al. 2012), and others [e.g.,

the Joint Effort for Data assimilation Integration (JEDI)] can greatly facilitate data assimilation for both research and operational use. In this paper, we introduce another such tool—version 2.0 of the open source statistical interpolation toolbox called Tendral Statistical Interpolation System (T-SIS).¹ The T-SIS version 2 presented here is significantly improved over T-SIS 1.0 (Halliwell et al. 2014) and can now handle additional algorithmic intricacies and complexities of observational data and models required to produce an integrated analysis.

The general problem of fusing measurements with a prior or background field leads to an overdetermined system of equations which is usually solved by imposing an optimality condition such as the least squares criterion. The solution is obtained either as an optimal estimator in a probabilistic framework or equivalently the combined errors of the overdetermined system are cast in quadratic form and standard convex optimization techniques are used to find the optimal solution (Crassidis and Junkins 2004; Kalnay 2002; Wunsch 1996). In both approaches, the solution procedure involves mathematical operations such as inversions and decomposition and factorization of the error covariance matrix. Furthermore, for an optimal estimation, the uncertainty in the state encoded by the covariance matrix has to be propagated through the process model which is usually nonlinear in the case of ocean models. The uncertainty covariance matrix is impractically large for real problems, and thus, various approximations and parameterizations have been used for applications each emphasizing different scales of the error

 Denotes content that is immediately available upon publication as open access.

Corresponding author: Ashwanth Srinivasan, a.srinivasan@tendral.com

¹ Downloadable from www.tendral.com/tsis/.

process and/or computational efficiency (Oke et al. 2002; Cummings 2005; Counillon and Bertino 2009).

Earlier versions of the package (T-SIS 1.0) presented herein implemented a classic multivariate optimal interpolation approach in which the large global error covariance matrix is split into smaller submatrices. The submatrices are built from a relatively small subset of local observations making the matrix sizes small and inversions feasible. Two well-known drawbacks of this method, the potential loss of coherence between large and small scales of the analysis due to the partitioning the observations and the $O(n^3)$ computational scaling with the number of observations n prompted a search for alternative methods. One particular approach for efficiently solving a large analysis problem is by modeling the prior/background error process as a Gauss Markov random field (GMRF) which leads to a sparse inverse of the error covariance matrix often referred to as the information or precision matrix.² The sparse representation of the global error covariance matrix can result in significant savings in computational time and memory burden without loss of accuracy (Rue and Held 2005). The analysis using the information matrix addresses the drawbacks mentioned above, it uses all of the data simultaneously and has a favorable scaling with the number of observations, and in this respect is similar to 3D-Var type of analysis. This approach was introduced for ocean data assimilation by Chin et al. (1999, 2002) and was found to be comparable in performance to more widely used schemes in a series of twin experiments (Srinivasan et al. 2011). Despite this initial promise, the methodology has not seen much use in either ocean forecasting or hindcasting applications. Recently, however, this technique was implemented in T-SIS 2.0 and it is presently being used operationally for relatively fine horizontal-resolution hindcasting and forecasting applications.

The purpose of this paper is to systematically document the capabilities of the latest T-SIS 2.0 package with an emphasis on the information matrix based analysis scheme and, furthermore, to demonstrate its performance in a global hindcasting application. In the following, a brief background and technical details of the analysis methods implemented in the T-SIS package are first presented in sections 2 and 3. In section 4, the performance of the various covariance models are then compared in simple test problems. In section 5, practical implementation details, observation types, and quality control are discussed. In section 6, the performance of these choices are then illustrated with a global hindcasting application similar in complexity and scope to applications used in many research and operational centers. This is followed by a discussion in section 7 to conclude the paper.

2. Statistical interpolation³

We suppose that the equations for the state \mathbf{x}_t and measurement vectors \mathbf{y}_t be represented as

$$\mathbf{x}_t = \mathbf{f}_t(\mathbf{x}_{t-1}) + \boldsymbol{\epsilon}_t^{\text{model}}, \quad (1)$$

$$\mathbf{y}_t = \mathbf{H}\mathbf{x}_t + \boldsymbol{\epsilon}_t^{\text{data}}, \quad (2)$$

where the spatially discretized state vector contains all dynamically independent prognostic variables at time step t . The function \mathbf{f}_t is a representation of a computation of a single time step by the ocean model and $\boldsymbol{\epsilon}_t^{\text{model}}$ is the aggregate uncertainty (both physical and numerical) in the model. We also make the convenient (but optional) assumption that the operator \mathbf{H}_t relates the state \mathbf{x}_t to observations by linear combinations. The errors in the models and observations are as yet of unknown character. The actual observed values of the \mathbf{y}_t are then used to constrain the ocean models dynamic trajectory toward the observations. Adopting a classical least squares approach, a cost function \mathbf{J}_N is minimized with respect to the model states $\mathbf{x}_0 \dots \mathbf{x}_N$:

$$\mathbf{J}_N = \|\mathbf{x}_o - \mathbf{x}_o^d\|_{\mathbf{L}_o}^2 + \sum_{j=1}^N \left[\|\mathbf{x}_j - \mathbf{f}_j(\mathbf{x}_{j-1})\|_{\mathbf{M}_j}^2 + \|\mathbf{y}_j^d - \mathbf{H}_j\mathbf{x}_j\|_{\mathbf{N}_j}^2 \right]. \quad (3)$$

The positive definite matrices \mathbf{L} , \mathbf{M} , and \mathbf{N} are weighting matrices, which are quantitative measures of our belief in the initial condition, the dynamical model, and the observations (Jazwinski 1970). The indices j represent time and o denotes initial conditions, respectively. The least squares formulation above is time dependent in that the optimal values of the weighting matrices depend on the time index j in (3) above. In practice, the above equation is solved in a recursive manner; that is, \mathbf{J}_{N+1} is estimated in terms of the observations \mathbf{y}_{n+1} using the current observations as a forcing term. The time recursion can then be expressed as

$$\mathbf{x}_t^f = \mathbf{f}_t(\mathbf{x}_{t-1}^a), \quad (4)$$

$$\mathbf{x}_t^a = \mathbf{x}_t^f + \mathbf{K}_t(\mathbf{y}_t - \mathbf{H}_t\mathbf{x}_t^f), \quad (5)$$

where the superscript f denotes forecast values and the matrix \mathbf{K} is referred to as the filter gain. To define an optimal filter gain, the weighting matrices \mathbf{L} , \mathbf{M} , and \mathbf{N} are treated as inverses of covariance matrices \mathbf{P} , \mathbf{Q} , and \mathbf{R} of errors in initial condition, dynamical model, and observations, respectively, and each with well-defined statistics.

There are numerous ways to represent the solution to this problem, two solutions given by the normal equations for this problems are (Kalnay 2002)

$$\mathbf{K}_t = \mathbf{P}_t\mathbf{H}_t^T(\mathbf{H}_t\mathbf{P}_t\mathbf{H}_t^T + \mathbf{R}_t)^{-1}, \quad (6)$$

² The term information matrix is used in the signal/image processing community, while the term precision matrix is used in statistics. The information matrix terminology is used in this paper.

³ Statistical interpolation is very wide subject used across many different fields. Here it discussed within the limited context of assimilation of data into numerical ocean models.

$$\mathbf{K}_t = (\mathbf{P}_t^{-1} + \mathbf{H}_t^T \mathbf{R}_t^{-1} \mathbf{H}_t)^{-1} \mathbf{H}_t^T \mathbf{R}_t^{-1}. \tag{7}$$

The optimal gain requires the use of nonstationary error covariance matrix \mathbf{P}_t in the above equations. For linear systems with Gaussian statistics, the Kalman filter recursion of the error covariance matrix provides the statistically optimal weights for (3). For nonlinear systems such as ocean data assimilation problems considered here, two approximate forms of the Kalman filter are generally used. The first of these uses the extended Kalman filter approach by linearizing the nonlinear model. The error covariance, \mathbf{P}_{t-1}^a , is then propagated as

$$\mathbf{P}_t^f = \mathbf{F}_t \mathbf{P}_{t-1}^a \mathbf{F}_t^T + \mathbf{Q}_t, \tag{8}$$

$$\mathbf{P}_t^a = (\mathbf{I} - \mathbf{K}_t \mathbf{H}_t) \mathbf{P}_t^f (\mathbf{I} - \mathbf{K}_t \mathbf{H}_t)^T + \mathbf{K}_t \mathbf{R}_t \mathbf{K}_t^T, \tag{9}$$

where the matrix $\mathbf{F}_t = \partial \mathbf{f}_t / \partial \mathbf{x}(\mathbf{x}_{t-1}^q)$ is the linearized model about the most recent estimate (Crassidis and Junkins 2004) and \mathbf{Q}_t is the model error covariance matrix. Here only the second moment of the state probability distribution is propagated, which can lead to closure issues.

A far more popular methodology for nonlinear problems is the ensemble Kalman filter introduced by Evensen (1994, 2006) and its variants. The ensemble Kalman filter deals with nonlinearity during the forecast step by working with ensemble of randomly generated states. Thus, it makes no restrictive assumptions during the state uncertainty forecast step, but there is an implicit Gaussian assumption at the state update step. Despite this compromise, the ensemble Kalman filter works well for many applications as shown by Evensen (2003).

Regardless of the methodology used, time recursion of the forecast error covariance matrix can be computationally orders of magnitude more costly than integrating the numerical model itself. Therefore, for many applications, the time recursion of the forecast error covariance matrix is neglected and, instead, a stationary error covariance matrix is used. Even without the time recursion, the error covariance matrices appearing in the normal Eqs. (6) and (7) are still too large to operate upon directly and therefore have to be further simplified for practical use. These simplifications must generate a symmetric positive definite matrix, and additionally it is often required to have properties such as smoothness, e -folding and should enforce balance properties such as geostrophy. Many different approximations of the error covariance matrix are currently used in operational systems (Martin et al. 2015). The T-SIS 2.0 package implements some of these approximations and introduces an additional one based on GMRF models for the error process. These are described in the next section. We do not address time recursion of the error covariance in this paper but mention in passing that with the inclusion of the GMRF based parameterization, T-SIS 2.0 allows for efficient time recursion of the error covariance matrix using either of the methods described above.

3. Modeling of the forecast error covariance matrix

The T-SIS 1.0 package initially performed the analysis based on Eq. (6) using a classic parameterization of the forecast error covariance, which was then augmented with a capability to use covariances derived from a set of model states or a set of mode vectors. More recently, in T-SIS 2.0, an analysis, which uses Eq. (7) and requires the inverse of the covariance matrix, was implemented using the GMRF paradigm. In this section, we first briefly describe the classic methods of representing the forecast error covariance (also available in T-SIS 2.0), followed by a more detailed discussion of the information matrix based method since it is not as common as the other methods.

a. Covariance using Gaussian type correlation functions or model states

T-SIS 2.0 retains the classic approach wherein the background error covariance \mathbf{P}^f is parameterized as a product of background error variances and correlations, modeled using second-order autoregressive functions (SOAR):

$$\mathbf{P}^f = \mathbf{D}^{1/2} \mathbf{C} \mathbf{D}^{1/2}, \tag{10}$$

$$\mathbf{C}_h = (1 + s_h) \exp(-s_h), \tag{11}$$

$$\mathbf{C}_v = (1 + s_v) \exp(-s_v), \tag{12}$$

$$\mathbf{C} = \mathbf{C}_h * \mathbf{C}_v; \tag{13}$$

s_h and s_v are the normalized horizontal and vertical correlations scales, \mathbf{C} is a matrix of correlations, and \mathbf{D} is a diagonal matrix of variances (Cummings 2005; Daley 1991; Kalnay 2002).

Alternatively, instead of analytical correlation functions a set of model states can be used to estimate the background error covariance matrix as

$$\mathbf{P}^f = \frac{1}{M-1} \sum_{m=1}^M (\mathbf{x}_m^f - \bar{\mathbf{x}}^f) (\mathbf{x}_m^f - \bar{\mathbf{x}}^f)^T, \tag{14}$$

where \mathbf{x}_m^f , $m = 1, \dots, M$, are the M model states (Counillon and Bertino 2009).

The states used for this purpose can be sampled from a long integration of the free running model or from data assimilated model runs when available. Sampling from a set of data assimilated model states have been shown to provide better estimates of the background covariance (Brasseur 2006; Panteleev et al. 2015). Anomalies for a given analysis time are calculated with respect to a running mean in order to estimate the error in the ocean state at a given period of the year for a chosen forecast horizon allowing for some time dependency. In typical use, the anomaly calculation retains the frequencies corresponding to ocean structures which are to be corrected by the data assimilation. The calculation of anomalies is implemented as part of the package and allows on-the-fly computation of anomalies for a given analysis time from a set of time-indexed states.

To compute an analysis, Eq. (6) is applied for each grid point using a subset of observations that fall within a few correlation scales (localization radius). The large global error covariance matrix in 6 is thus split into many local ones to make the matrix inversion and multiplication manageable. Essentially, this algorithm requires the solution of an $n \times n$ linear system for which the operation count and memory requirements scale as $O(n^3)$ and $O(n^2)$, respectively. This implies that the above algorithm can be slow for large correlation scales or dense datasets. However, these computations can be done in parallel and good performance is obtained by computing the analysis for each grid point separately on multiple processing elements. The analysis with analytical covariance models is generally faster compared to analysis with covariance derived from a set of model states, particularly for larger model state sizes, due to the need to load, store and operate on several model states at once. For many applications, good improvements in performance can be obtained by replacing the model states by a smaller number of mode vectors or empirical orthogonal functions (EOFs) to reduce input/output (I/O) and memory requirements and by computing the analysis in ensemble space. The transformation to ensemble space replaces the solution of the $n \times n$ system with a $p \times p$ system, where p is number of states or mode vectors and is computationally faster when $p \ll n$. T-SIS 2.0 can compute the analysis in both observation and ensemble space and the choice is usually application dependent.

b. Information matrix using Gaussian Markov random fields

The aforementioned computational bottleneck can be greatly improved by replacing the covariance matrix by a sparse approximation of its inverse, the information matrix. Such a compact representation of the information matrix can be realized by spatial regression (Chin et al. 1999). This approach has its roots in the concept of conditional autoregression introduced by Besag (1974) for defining statistical models of a spatially distributed random vector. Each element of the random process is specified conditionally on values of a few neighbors and such a process is known as a Markov random field (MRF). Formally, an MRF random vector $\mathbf{x} = (x_1, \dots, x_n)^T$ defined over a set of discrete locations such as a grid or lattice has a Markov property: an element x_i at a grid location I is conditionally independent from all other elements given the values of neighbors of i . This set of neighbors, or *neighborhood*, formalizes the notion of “locality” by specifying the extent of direct interactions among the grid elements (e.g., discretized forms of partial differential “balance” equations). This notion of locality is in contrast to the “localization” procedure in section 3a or as in Anderson (2012) that removes long-distance covariances in order to make the analysis methods practical. The information matrix is local, but its inverse the covariance matrix is not necessarily local and preserves and smoothly tapers long distance correlations (Rue and Tjelmeland 2002).

The specific form of MRF relevant here is the GMRF, whose random vector has a jointly Gaussian distribution, and

the conditional distribution of x_i given the values of its neighbors is

$$x_i \sim \mathcal{N}\left(\sum_{j \in \delta_i} \alpha_{ij} x_j, \beta_i^2\right), \tag{15}$$

where δ_i is the set of neighbors for particular location i . (Conventionally, the neighbor set δ_i does not include i .) Modeling with a GMRF consists of specifying the neighborhood system δ_i with its interaction parameters α_{ij} and the error variances β_i^2 . It can be shown (Rue and Held 2005) that Eq. (15) holds if and only if the joint distribution of \mathbf{x} is Gaussian and of the form

$$\pi(\mathbf{x}) = (2\pi)^{-n/2} |\mathbf{L}|^{1/2} \exp\left\{-\frac{1}{2} \mathbf{x}^T \mathbf{L} \mathbf{x}\right\}, \tag{16}$$

where \mathbf{L} is the information matrix (inverse of the covariance matrix)

$$\mathbf{L}_{ij} = \begin{cases} \frac{1}{\beta_i^2} & \text{if } j = i \\ \alpha_{ij} & \text{if } j \in \delta_i \\ 0 & \text{otherwise} \end{cases}. \tag{17}$$

The conditional distribution of x_i in Eq. (15) has the form of a Gaussian linear regression, in which the extent (“cardinality”) of the neighbor set δ_i determines the order of this regression. The cardinality of the neighborhood in GMRF also determines the sparsity of the information matrix \mathbf{L} whose diagonal elements are the conditional precisions $1/\beta_i^2$ and off-diagonal elements are the regression parameters α_{ij} . The regression parameters encode the correlations spanning all distances and represent the covariance matrix of \mathbf{x} in a numerically efficient, sparse form. For example, if the interaction among the elements of \mathbf{x} can be modeled with a system of first-order partial difference equations (over a two- or three-dimensional grid), then the neighborhood would consist only of the direct/nearest neighbors and the resulting \mathbf{L} is a nested tridiagonal matrix. Such sparseness in the information matrix enables significant computational advantages.

To determine the elements of the information matrix, a zero-mean process with identity covariance matrix is first defined as $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ and then \mathbf{x} is obtained through the inverse transform $\mathbf{z} = \mathbf{\Gamma} \mathbf{x}$. When \mathbf{x} is Gaussian with zero mean, the covariance is $\mathbf{P} = (\mathbf{\Gamma}^T \mathbf{\Gamma})^{-1}$, where $\mathbf{\Gamma}$ is the square root information matrix. A simple example is the AR(1) model on a regular line for which $\mathbf{\Gamma}$ and the corresponding information matrix \mathbf{L} are of the form:

$$\mathbf{\Gamma} = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -1 & -1 \end{bmatrix} \quad \mathbf{L} = \begin{bmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 2 \end{bmatrix}. \tag{18}$$

Another insightful view of the GMRF is as a discretized solution of stochastic differential equation. Consider the equation $\partial x/\partial s = \delta$, where δ is a white noise process corresponding

to the AR(1) model above. Using forward differences for regularly spaced locations ($\Delta s = s_{i+1} - s_i$), $\partial x / \partial s = [x(s_{i+1}) - x(s_i)] / \Delta s$. The discretization implies that $\delta s^{-1} \Gamma \mathbf{x} \sim \mathcal{N}(0, \delta s^{-2} \mathbf{I})$, which has a density proportional to $\exp[-1/(2\Delta s^2)](\Gamma \mathbf{x})^T (\Gamma \mathbf{x})$, or that $\mathbf{x} \sim \mathcal{N}(0, s^{-2} \mathbf{L})$. The one-dimensional example above can be extended to grids of arbitrary dimensions by extending the definition of Γ to contain one row for every pair of adjacent nodes in a d-dimensional grid. In the case of uniform grids, this leads to a simple structure for $\Gamma = \Gamma^T \Gamma$, namely,

$$\mathbf{L}_{i,j} = \begin{cases} |n_i| & \text{if } i = j \\ -1 & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases}, \quad (19)$$

where $|n_i|$ is the number of neighbors of a node i , for example, $|n_i| = 4$ for nonboundary nodes of 2D and $|n_i| = 6$ in 3D and so on. Higher-order GMRFs can be constructed by considering higher-order differences in the definition of the Γ operator. The elements for the second- and higher-order GMRFs are found by convolving the first-order GMRF with itself (Lindgren et al. 2011). In both cases, the information matrix is sparse and with $n + 1 + N$ (homogeneous case) or nN (inhomogeneous) elements replacing the N^2 elements in the covariance matrix, where n is the size of the numerical Laplacian or biharmonic stencil.

For our purposes the computational grid used by the ocean model defines a set of nodes i , following Chin et al. (1999), we write the elements of the error vector using generalized regression as

$$e_j = \sum_{i \in Z} \alpha_{ij} e_{j-1} + v_j, \quad (20)$$

where α_{ij} is inhomogeneous, Z a multidimensional, noncausal index set used to capture conditional dependence, and v_j are zero-mean Gaussian distributed random errors with variance β_j^2 . These spatial regression coefficients can be related to the inverse of the covariance matrix or the information matrix by writing the above equation as a matrix operator Γ_k on the error vector:

$$\Gamma_k e_k = \delta, \quad (21)$$

where δ denotes a zero mean process with identity matrix as covariance. Γ_k is sparse with the diagonal elements given by $1/\beta_j$ and near diagonal elements $-\alpha_{ij}/\beta_j$, $i \in Z$ for each row j . Sparseness of Γ depends on the size of the index set Z . Multivariate relationships between geopotential and velocity components can be encoded by the regression operator and the corresponding information matrix can be obtained by squaring the regression operator as above (Chin et al. 1999). Once the \mathbf{L} matrix is available, the optimal filter gain in Eq. (7) can then be obtained as $\mathbf{L}_a^{-1} \mathbf{H} \mathbf{N}$ where $\mathbf{L}_a = \mathbf{L}_f + \mathbf{H}^T \mathbf{N} \mathbf{H}$ and $\mathbf{N} = \mathbf{R}^{-1}$. In practice, the gain is not explicitly computed but $\mathbf{H} \mathbf{N} [\mathbf{y} - \mathbf{H}(\mathbf{x})]$ is first computed and then inverted with the sparse \mathbf{L}_a usually with 20 iterations of a preconditioned conjugate gradient algorithm. This algorithm has linear scaling with the analysis grid size.

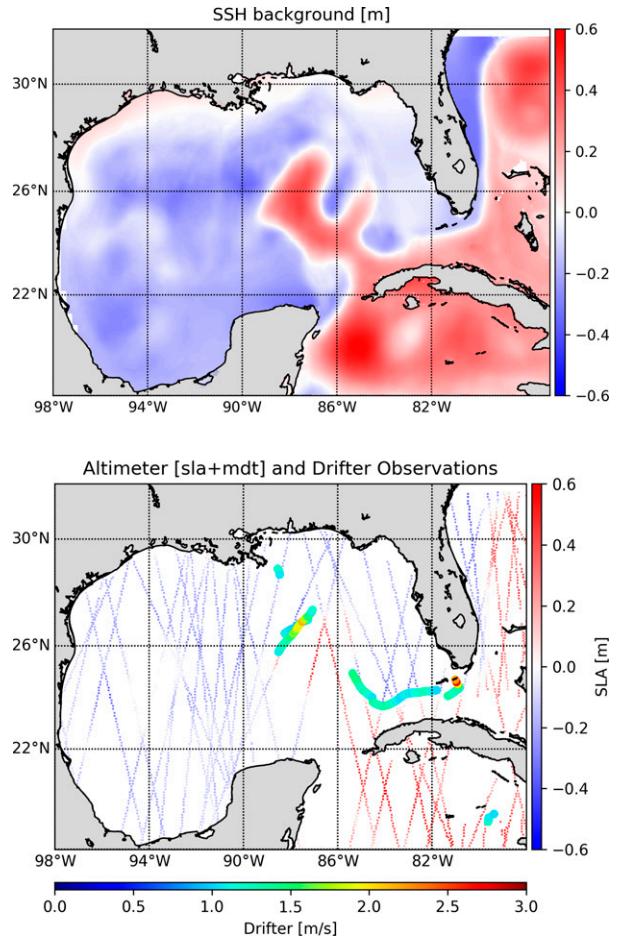


FIG. 1. (top) The background SSH field, (bottom) along-track SLA observations, and drifter velocities used to infer corrections to the background.

4. Examples of analysis with covariance and information matrices

In this section, we provide some examples of analysis computed using covariance and information matrices with Eqs. (6) and (7), respectively. First, we use along-track sea level anomaly (SLA) data from altimeters to estimate corrections to the background sea surface height (SSH) field from a $1/16^\circ$ model of the Gulf of Mexico (Fig. 1). The altimeter signal in the Gulf of Mexico is quite strong due to the Loop Current and is thus well suited for this illustrative example (Srinivasan et al. 2011). The ratio of the observations to background variances is set to 5 in the examples shown in Fig. 2.

Two different analyses are computed for this illustration using both covariance and information matrix methods (Fig. 2). The analysis based on the covariance matrix and Eq. (6) were computed using two different correlation scales, one of 150 km and the other of 300 km corresponding to roughly 3 and 5 times the Rossby deformation radius of the first baroclinic mode. For the information matrix based method, we use both first- and second-order information matrices as defined in the previous section. The first-order model uses four nearest

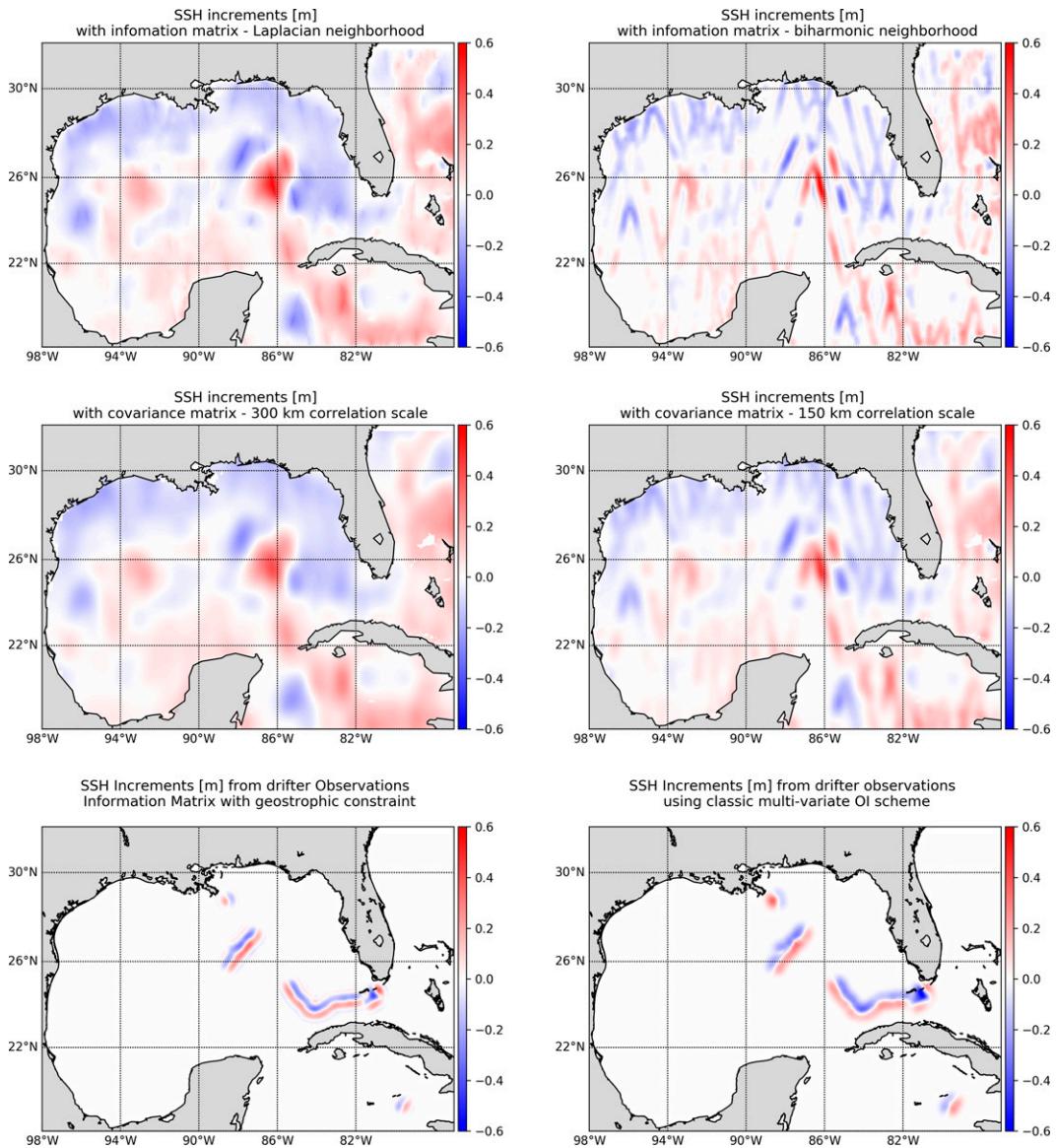


FIG. 2. (top),(middle) Corrections to the SSH background estimated using Eqs. (6) and (7). Two different neighborhood systems (Laplacian and biharmonic stencils) were used in the information matrix in the top row. For comparison, corrections estimated using SOAR functions with correlation scales (localization radius) of 300 and 150 km are shown in the middle row. (bottom) Multivariate surface height corrections estimated from the drifter observations in Fig. 1. (bottom left) Corrections derived from the information matrix can be compared with (bottom right) corrections from a classic multivariate OI. The background state velocity was assumed to be zero over the domain.

neighbors in the grid along cardinal directions corresponding to a Laplacian stencil while the second-order model uses 12 neighbors and corresponds to a biharmonic stencil. First- and second-order \mathbf{L} take the form

$$\mathbf{L}_1 = \begin{bmatrix} & -1 & & & \\ -1 & 4 & -1 & & \\ & -1 & & & \end{bmatrix} \quad \mathbf{L}_2 = \begin{bmatrix} & & 1 & & & \\ & 2 & -8 & 2 & & \\ 1 & -8 & 20 & -8 & 1 & \\ & 2 & -8 & 2 & & \\ & & & & & 1 \end{bmatrix}$$

The corrections computed using the first-order information matrix are both smoother and extend further than the corrections

with the second order (top two rows of first column, Fig. 2). The first-order GMRF (Laplacian stencil) is built with first-order increments (section 3b) while second-order (biharmonic stencil) is built with second-order increments. The error vector which is written as a regression is expected to be smoother with biharmonic stencil (more neighbors) than a Laplacian stencil (less neighbors). However, since the analysis is computed by inverting the information matrix we get smoother results for the first-order GMRF. This can be counter intuitive when looking at it from first-order versus second-order GMRF perspective. However, the information matrix is a “roughening” operator and its inverse

TABLE 1. Sample analysis times for different schemes as implemented in the T-SIS 2.0.

Method	Observation count	Grid size	Time
Covariance matrix; Eq. (6)	200 000	1500 × 1100	4 min
Information matrix; Eq. (7)	200 000	1500 × 1100	45 s
Covariance matrix; Eq. (6)	1 000 000	1500 × 1100	30 min
Information matrix; Eq. (7)	1 000 000	1500 × 1100	1 min

will act as a smoothing operator. Lindgren et al. (2011) show that the first-order model approximates a Gaussian process with covariance, $\text{cov}(x_{ij}, x_{i'j'}) = K_0(kr)$, where K_0 is the modified Bessel function of the second kind, k is a constant and r is the distance between x_{ij} and $x_{i'j'}$ with the spatial range parameter $\rho \rightarrow \infty$. The second-order information matrix corresponds to the widely used second-order autoregressive model with covariance, $krK_1(kr)$, equivalent to a thin-plate spline (Rue and Held 2005). Paciorek (2013) compared the eigenstructure and smoothing kernels of both the first- and second-order models above and finds that the first-order model spreads the corrections farther than the second-order model similar to the results presented here. The MRF model structure and parameterization are fundamental, but errors or inaccuracies in the iterative inversion of the information matrix can also play a significant practical role. For example, incomplete iterations could yield smoother fields than the optimal solution. Therefore, the solutions were examined for both the standard 20 iterations and also for 100 iterations. There were no significant difference in this case. For comparison, the corrections computed using SOAR functions with a large correlation scale produces a smoother estimate than the one with the small correlation scale as more observations are used for each grid point (middle row, Fig. 2). Interestingly, these corrections (with correlation scales of 150 and 300 km) are very similar to the corrections computed with the biharmonic and Laplacian stencil based information matrices suggesting that the information matrix defined by strictly local differential operators represents covariance structure over a much larger region in space efficiently.

Multivariate corrections for surface height computed from drifter velocity observations are shown in the bottom row. A simple geostrophic balancing constraint written in the form of a regression operator is used to define the information matrix for this analysis (Chin et al. 1999). As seen in Fig. 2, this approach results in an updates which are almost identical to ones calculated using classic multivariate optimal interpolation schemes as in Daley (1991) or Cummings (2005). Only velocity observations were used to compute surface height increments in this illustrative example. In typical usage, all of these different observations are combined for a joint estimate.

Overall, the corrections computed for the two different cases and analysis methods are, as expected, very similar and suggest that either of these methods can be used with similar impact for actual applications. However, there are some clear advantages to using the information matrix method and Eq. (7). First, a global analysis is computed and there is no need for partitioning the data into subsets which makes it possible to maintain coherence between large and small scales in the analysis without resorting explicit smoothing. Second, the

analysis computation and storage cost scale linearly with the analysis grid and observation size compared to the cubic and quadratic increase for the implementations using Eq. (6) (Table 1). Furthermore, the commonly used Laplacian and biharmonic stencils require matrix-vector product with at most a 5×5 footprint, which is usually less than the buffer or halo regions for tiles in domain decomposed ocean models. Therefore, the GMRF based analysis can be implemented as a subroutine within the ocean model with relatively less effort and thereby making long hindcasts more efficient and opens up the possibility of time recursion of the information matrix through the numerical model.

5. Practical implementation

The theoretical basis of the assimilation scheme used to produce the spatial analysis was described and illustrated in the previous sections. In this section, we provide some details on the practical implementation of the T-SIS 2.0 ocean data assimilation package.

a. Error variances

All schemes implemented in the package require the specification of background/forecast and observation error variances which vary with analysis variable, observation types, depth, and location. In the absence of a time history of forecast errors or statistics of innovations or increments, background error variances are specified using climatology. Climatological variability either from model runs or gridded climatology based on observations can be interpolated in space and time and scaled by a factor and specified as background variance. Climatological variability derived from data assimilative model run is provided as part of the package for this purpose. This can then be replaced by actual error statistics once sufficient time history of errors are built up using one of the many methods in the literature (e.g., Hollingsworth and Lönnberg 1986). Error covariances are derived directly from the ensemble of model states when the ensemble version is used. These covariances are scaled to reduce the magnitude of the error variances as errors of the days are expected to be generally smaller than the climatological variability. The observation error covariance matrix is a sum of measurement error and representativity error. Representativity errors are implemented here as a function of the model resolution and observation networks. The observation errors are usually assumed to be uncorrelated.

b. Observation preprocessing and quality control

Any variable directly related to the state vector (velocity components, surface height, temperature, salinity) can be

TABLE 2. Ocean observations that have been used with TSIS 2.0.

No.	Observation	Provider/source	Type
1	Sea level anomalies (SLA)	CLS marine.copernicus.eu	Along-track and gridded maps
2	Foundation temperature/SST	OISST (NOAA) and K10-SST (NAVO)	Gridded L4 products
3	In situ SST	NOAA-IQUAM	Point
4	In situ	GTS	Point
5	Argo <i>T/S</i>	U.S. GODAE	Point
6	XBT	GTS and offshore industry	Point
7	<i>T/S</i> from moored buoys	GTS	Point
8	Surface salinity maps	marine.copernicus.eu	Gridded L4 products
9	Drifter velocities	GDP and offshore industry	Point
10	HF-radar velocities	NOAA and offshore industry	Gridded maps
11	Vessel mounted ADCP data	Offshore industry	Velocity profiles
12	ROCIS surface velocity data	Offshore industry	Surface velocity along flight tracks

assimilated relatively easily using T-SIS 2.0. At present, the package can assimilate a variety of remotely sensed and in situ observations available from different sources (Table 2). Along-track SLA from multiple altimeters can be pooled and assimilated. Alternatively, maps of SLA can also be assimilated. In all cases the anomalies are converted into a surface height by adding a mean dynamic topography usually derived from long model simulations. There are numerous sources of sea surface temperature (SST) data each with its own biases and errors. For many applications, gridded fields of SST can simplify the need to keep track of errors and biases of each independent source. Two such products, the 1/4° optimally interpolated SST from National Oceanic and Atmospheric Administration (NOAA) and the 10 km product from the Naval Oceanographic Office (NAVO) are supported as default options in the package. Additional products can be easily added as required. Alternatively, composite satellite retrievals with error estimates (Piolle et al. 2010) project and in situ SST observations (Xu and Ignatov 2014) can be used for SST assimilation. Temperature and salinity profiles from sources in Table 2 can be assimilated. Support programs to convert profiles to various vertical coordinates, to extend profiles from the last observed level to the bottom, to resample the vertical profiles and statistical regression coefficients to estimate salinity from observed temperature profiles are available as part of the package. Velocity data from sources in (Table 2) can also be assimilated.

Before assimilation, all observations that are available for analysis for a particular time/date are preprocessed and collected in an intermediate observations file. Assimilated observations are typically obtained from quality controlled sources for most applications other than operational applications using raw data from the Global Telecommunication System (GTS). Even so, all observations undergo a simple quality control procedure before being included in the observations file. These include gross error checks such as physical range checks, land-sea mismatch checks, data-time checks, and checks against climatology and forecast fields. Observations differing by more than a user specified factor from the climatological standard deviation are either not assimilated or moderated.

c. Vertical coordinates for analysis and analysis variables

The package was originally developed for use with the Hybrid Coordinate Ocean Model (HYCOM) and supports direct analysis in generalized vertical coordinates. The state vector used is a subset of the prognostic variables of HYCOM, specifically layer thickness, layer temperature, layer salinity, the diagnosed SSH anomaly, and the horizontal velocity components. To optimize system performance for the Lagrangian vertical coordinate system, observations are first remapped onto the hybrid isopycnal-sigma- z vertical coordinate system prior to assimilation. The analysis procedure then updates each layer independently in a vertically decoupled manner. For temperature-only profiles, corresponding salinity profiles are generated from climatological temperature-salinity (T - S) relationships to permit layerization (Thacker 2008). In the pressure layers both temperature and salinity are corrected. In isopycnal layers, thickness either temperature or salinity is corrected with the other diagnosed from the equation of state. Profile data are first assimilated to obtain an intermediate state which is then corrected with altimeter data. A layerized version of the Cooper and Haines (1996) procedure (described in the appendix) is used to adjust the layer thicknesses in the isopycnal-coordinate interior in response to SSH anomaly innovations. Multivariate correlations are used to compute velocity corrections from layer thickness corrections and vice versa.

Alternatively, the analysis can be computed in standard physical coordinates with pressure as the vertical coordinate. In this case, the state vector consists of temperature, salinity and dynamic height computed from temperature and salinity. All variables are analyzed simultaneously in three dimensions. The analysis in layered coordinates is computationally faster but the quality of the analysis appears to be similar in limited tests. Further systematic comparisons are required to fully evaluate these options.

6. An example hindcast application

We proceed now to illustrate the performance of the information matrix based scheme with a global 1/4° hindcast application with HYCOM.

TABLE 3. Observations for assimilation and verification for the 2018 hindcast.

No.	Observation	Assimilation	Verification
1	SLA	Along-track data	1/4° gridded SLA maps from AVISO
2	Foundation temperature/SST	NOAA 1/4° OISST	1/4° gridded GMPE SST
3	<i>T/S</i> profiles	Argo <i>T/S</i> profiles	Withheld <i>T/S</i> from moored buoys
4	Velocities	Not assimilated	GDP kriged drifting buoy velocities

a. System configuration

1) THE 1/4° GLOBAL HYCOM

The HYCOM model, its capabilities and reference simulations are detailed in (Bleck 2002; Chassignet et al. 2003; Halliwell 2004). The implementation of HYCOM used here is similar to configurations used in other HYCOM based operational centers such as Naval Research Laboratory and National Centers for Environmental Prediction (Chassignet et al. 2009). For this global application, a composite horizontal grid consisting of a uniform latitude–longitude grid between 78° and 60°S, a Mercator grid between 60°S and 47°N with a 1/4° grid spacing, and a curvilinear rotated pole grid north of 47°N is used. The resolution at the equator is approximately 28 km and reduces to less than 10 km at high latitudes. The model bathymetry is a combination of global 1-arc-min ocean depth and land elevation from the U.S. National Geophysical Data Center (ETOPO1; <https://www.ngdc.noaa.gov>) and the 15-arc-s General Bathymetric Chart of the Oceans (GEBCO; https://www.gebco.net/data_and_products/gridded_bathymetry_data/) 2020 grid with local corrections in the Indian Ocean, Gulf of Mexico, and Brazil Current regions.

In the vertical, the model is configured with 32 hybrid (pressure–sigma–isopycnal) layers with potential densities referenced to 2000 db and ranging from 27.10 to 37.17 sigma units. Surface atmospheric forcing is derived from the ECMWF reanalysis v5 (ERA5) dataset (Hersbach et al. 2018) and consists of three hourly fields of air temperature and specific humidity at 2 m, surface net downward and long- and short-wave radiation, precipitation, and 10 m wind speeds. The atmospheric radiative fluxes are scaled using CERES energy balanced and filled gridded product (Kato et al. 2018). Momentum and heat turbulent surface fluxes are computed from CORE 2 bulk formulas (Large and Yeager 2004). Monthly climatological river discharge is used to specify a virtual salinity flux to include the effects of river inflow. Surface salinity is relaxed to the *World Ocean Atlas 2018* salinity with an *e*-folding time of 1 month. A combination of Laplacian and biharmonic mixing is used for horizontal momentum diffusion while a biharmonic formulation is used for horizontal thickness diffusion. These are specified with diffusion velocities of magnitude 0.003 m s⁻¹ for the Laplacian term and 0.02 m s⁻¹ for the biharmonic mixing terms, respectively. The *K*-profile parameterization (KPP; Large et al. 1994) is used for vertical mixing with default values. Finally, a simple thermodynamic energy-loan model is used for heat balance in regions with ice.

2) OBSERVATIONS

Observations used for assimilation and verification are listed in Table 3. Remotely sensed along-track sea level anomalies

(SLA) from Collecte Localisation Satellites (CLS), gridded maps of 1/4° Optimally Interpolated Sea Surface Temperature (OISST) from NOAA and in situ temperature/salinity (*T/S*) profiles from the Argo program obtained from usgodae.org are systematically assimilated daily during the course of the hindcast experiment (Fig. 3).

Along-track SLA data of 7-km nominal resolution from six altimeters, *Jason-3*, *CryoSat*, *Sentinel-3A* and *Sentinel-3B*, *AltiKa*, and *HY-2B*, are sampled from a 7-day window (± 3 days) centered on the analysis data are pooled together and assimilated daily (data coverage for a typical day from these altimeters is shown in Fig. 3a). A reference mean dynamic topography (MDT) based on Centre National d'Études Spatiales–Collecte Localisation Satellites 18 MDT (Mulet et al. 2021) is added to the anomalies to convert the anomaly fields into the SSH fields. For the altimeters used in this experiment, the data provider, CLS, suggests instrument errors ranging from 2 to 4 cm. However, we used a constant 7-cm error for the altimeter data; this is slightly on the higher side than instrument accuracy but compensates for unknown errors in MDT and representativity errors in a crude way for this demonstration experiment. The observation errors are inflated to give more weight to the observations closest to the analysis day with a Gaussian weighting scheme, which smoothly sets the errors after 10 days to climatological levels.

Daily 1/4° OISST analysis and the associated formal mapping errors were bilinearly interpolated to the model grid for assimilation. Typical SST errors range from 0.2° to 1.5°C and are usually high in regions of high mesoscale activity.

Delayed mode Argo profiles are used for this hindcast. Globally, there are about 250–400 profiles available each day. As for the altimeters, profiles falling within a time window of 7 days (± 3 days) centered on the analysis date are used for assimilation. Temperature and salinity error standard deviations were specified as 0.05°C and 0.02, respectively, and were further adjusted to give more weight to the observations closest to the analysis day.

Both observation based gridded products and independent withheld observations are used for verification. For SLA, optimally interpolated maps at 1/4° produced daily by Archiving, Validation, and Interpretation of Satellite Oceanographic Data (AVISO; Le Traon et al. 1998) by merging data from all available altimeters are used to assess the consistency of hindcast. Similarly, for SST the Met Office provides a 1/4° Global Median Product Ensemble (GMPE) SST product from several daily SST analysis produced by the Global High-Resolution Sea Surface Temperature (GHRSSST) consortium. The resolution of both these maps (1/4°) is the close to the model's native resolution making it particularly convenient for assessing the

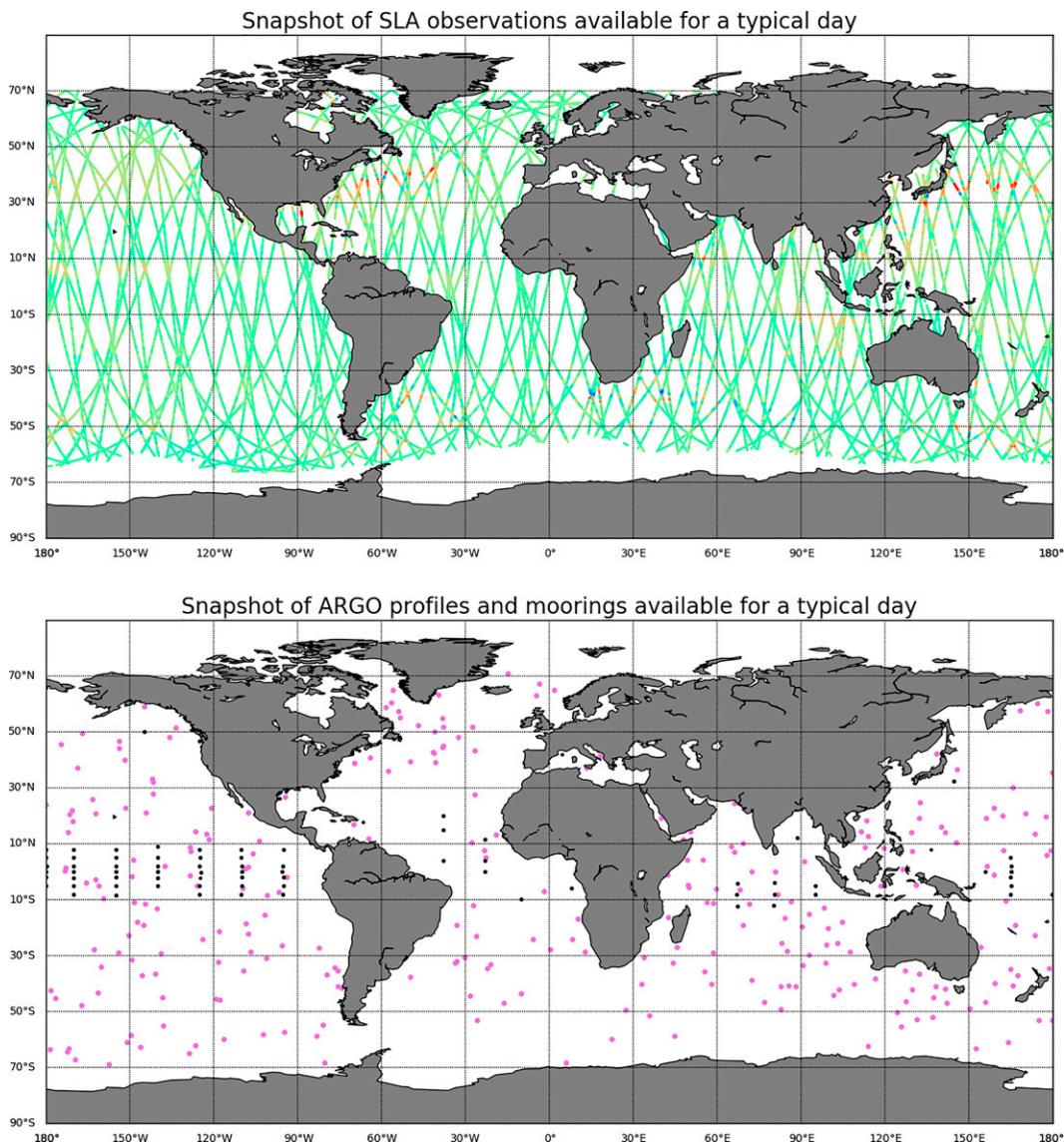


FIG. 3. A sample of the observations available on any given day. (top) Remotely sensed along-track sea level anomalies (SLA) from multiple altimeters. (bottom) In situ temperature/salinity (T/S) profiles from the Argo program in magenta color and mooring locations are shown in black.

hindcast. We use both the assimilated Argo observations and withheld T/S data from the tropical moored buoys to evaluate the analysis. Finally, we use withheld velocity data (Lumpkin and Centurioni 2019) from the NOAA Global Drifter Program (GDP; <https://www.aoml.noaa.gov/phod/gdp/index.php>) to assess the 15-m velocity fields produced by the hindcast.

3) DATA ASSIMILATION, INITIAL CONDITIONS, AND THE ANALYSIS-UPDATE CYCLE

The state vector used in this hindcast is a subset of the prognostic variables of HYCOM, namely, baroclinic layer thickness, layer potential temperature, layer salinity, and layer velocities.

The spatial analysis is computed using the information matrix with a biharmonic stencil as detailed in section 3b. Spatially varying background error variances are specified using climatological values (scaled by a factor of 0.6) derived from a prior data assimilative experiment. T/S /layer thickness data from layerized profiles are first assimilated to produce an intermediate analysis. The sea surface height computed from this intermediate analysis is then used a background field for a second and final SSH analysis, this time using along-track altimeter data. This final analyzed “SSH” is then used to further adjust the water column density field with the “layerized” Cooper and Haines (1996) algorithm described in the appendix. Here the SSH innovations are transformed into layer thickness adjustments. In

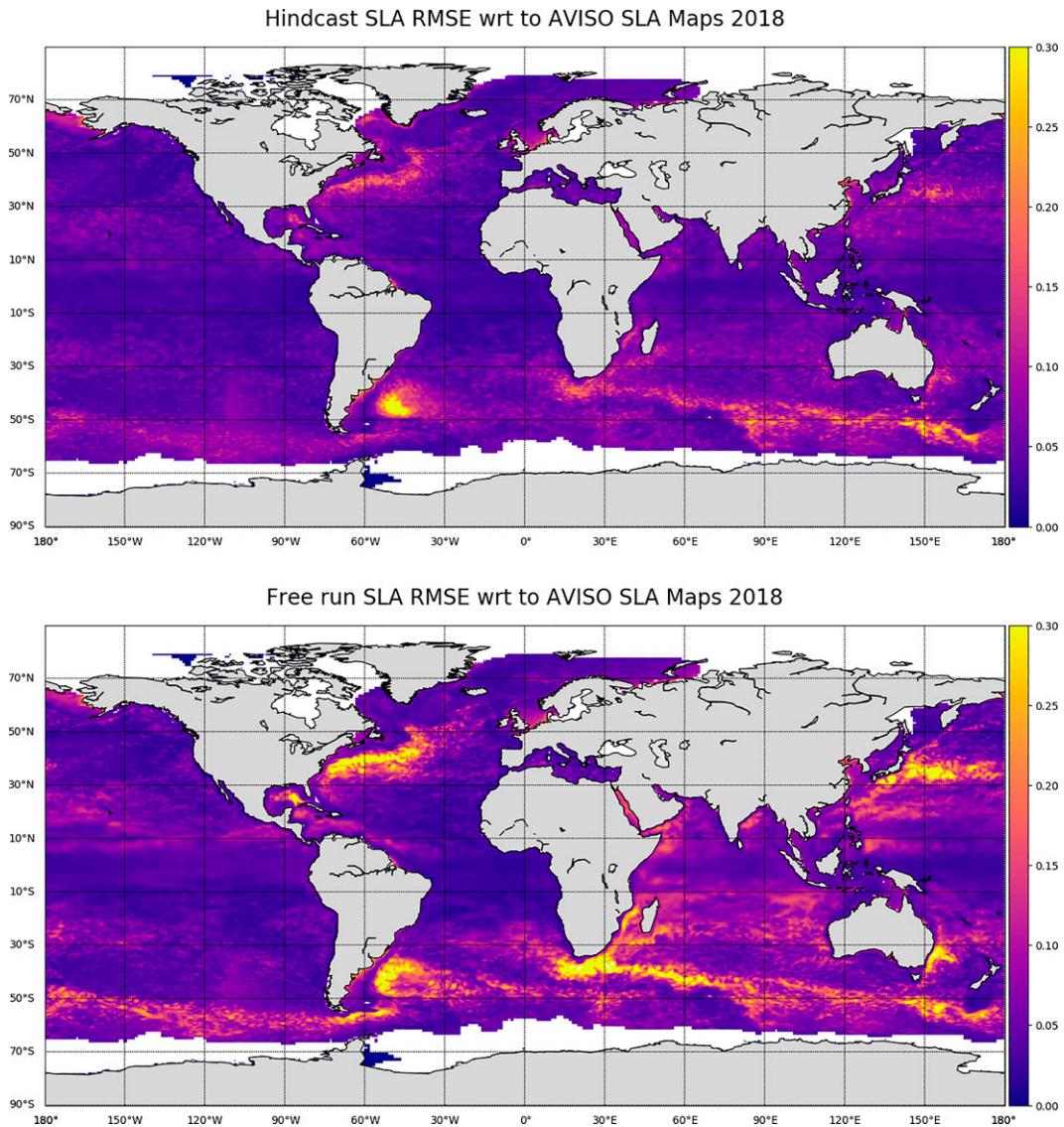


FIG. 4. RMS errors in SLA computed with respect to the AVISO gridded SLA. (top) RMS error from the hindcast and (bottom) RMS error from the free run. The spatial distribution of the RMS errors is calculated for each grid point with data for the duration of the hindcast.

doing so, care is taken to avoid using this algorithm to within one correlation scale of actual profiles to ensure that corrections to the water column using actual T/S profiles are not overwritten by the altimeter derived adjustments. The analysis produces increments to temperature, salinity, layer thickness, and horizontal velocity components. However, since there are no velocity observations, multivariate velocity corrections are not explicitly used instead the model velocity is allowed to adjust to the pressure increments over a 6-h time frame.

For this hindcast, initial conditions for 1 January 2018 is a nowcast produced by an operational system using the older T-SIS 1.0 version. Starting from this initial conditions the model was advanced daily to 1800 UTC and an analysis is computed using observations valid within a ± 3 -day window

and centered at 0000 UTC the following day. The prognostic variables are then incrementally updated (Bloom et al. 1996) using constant increments over a 6-h period, starting at 1800 UTC and produce a nowcast or analysis at 0000 UTC the following day. These nowcasts are saved for the system evaluation presented here. The same initial conditions were used for a companion free run and advanced for the entire 2018 year with no data constraints.

b. System evaluation

We begin by examining the impact of altimetry. Instead of directly comparing with along-track data used in the assimilation process, we present comparison with gridded maps produced by

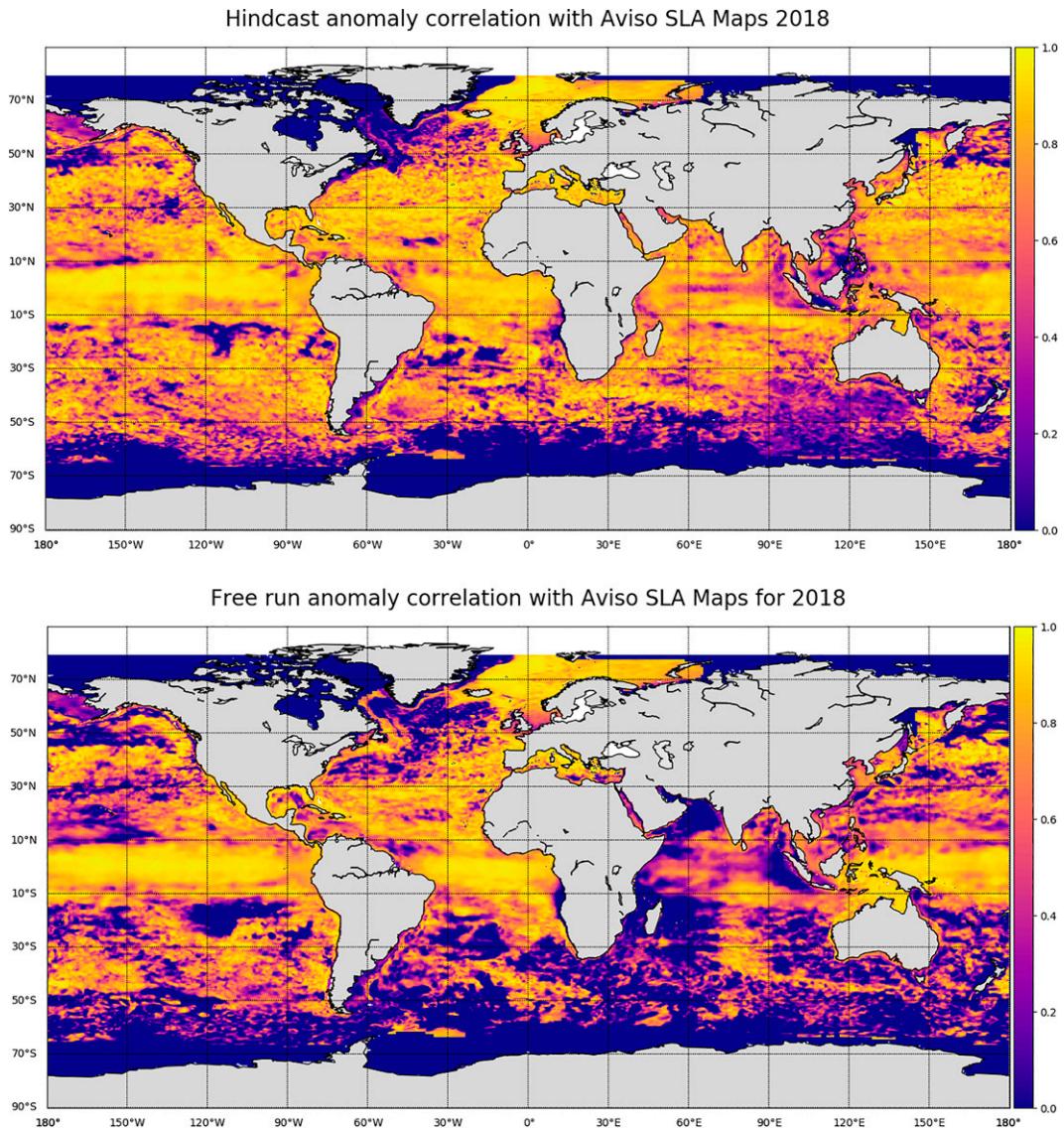


FIG. 5. SLA anomaly correlation computed with respect to the AVISO gridded SLA maps. The correlation between AVISO and model SLA for (top) the hindcast and (bottom) the free runs. The correlations were computed in time for each grid point.

AVISO. Maps of RMSD for the hindcast and the free run with respect to the AVISO gridded product (Fig. 4) shows that the RMSD for the hindcast over most of the ocean is close to the specified SLA error of 7 cm and is everywhere lower than the free run. The error levels in the assimilated product are generally less than half the magnitude of the errors in the free running simulation and well below the observation standard deviation. However, in areas of high mesoscale activity such as the Gulf Stream, Kuroshio, Agulhas, the RMSD while still lower compared to the observed standard deviation, is higher than the specified error suggesting room for improvement. A spatial map of the temporal correlation between AVISO maps and instantaneous modeled fields (Fig. 5) shows a good anomaly correlation (>0.75) in most regions where SLA data are assimilated. The free run SLA

exhibits relatively weaker correlation, particularly in the Southern Hemisphere, which is rectified in the hindcast. Apart from the areas of strong mesoscale activity, the RMS errors increase at high latitude in the circumpolar Southern Ocean. A combination of mismatch in correlation scales due to the uniform grid used in this region and aliasing of the fast barotropic signals due to strong winds in this regions are likely reasons for this increased errors.

Both AVISO maps and hindcast use the same set of input SLA data but they use different procedures for the estimates, and thus, the above comparison can be considered as a necessary test for consistency. The AVISO product uses a much longer time window (35–42 days; <https://catalogue.marine.copernicus.eu/documents/QUID/CMEMS-SL-QUID-008-032-062.pdf>) of observations to compute the gridded estimates

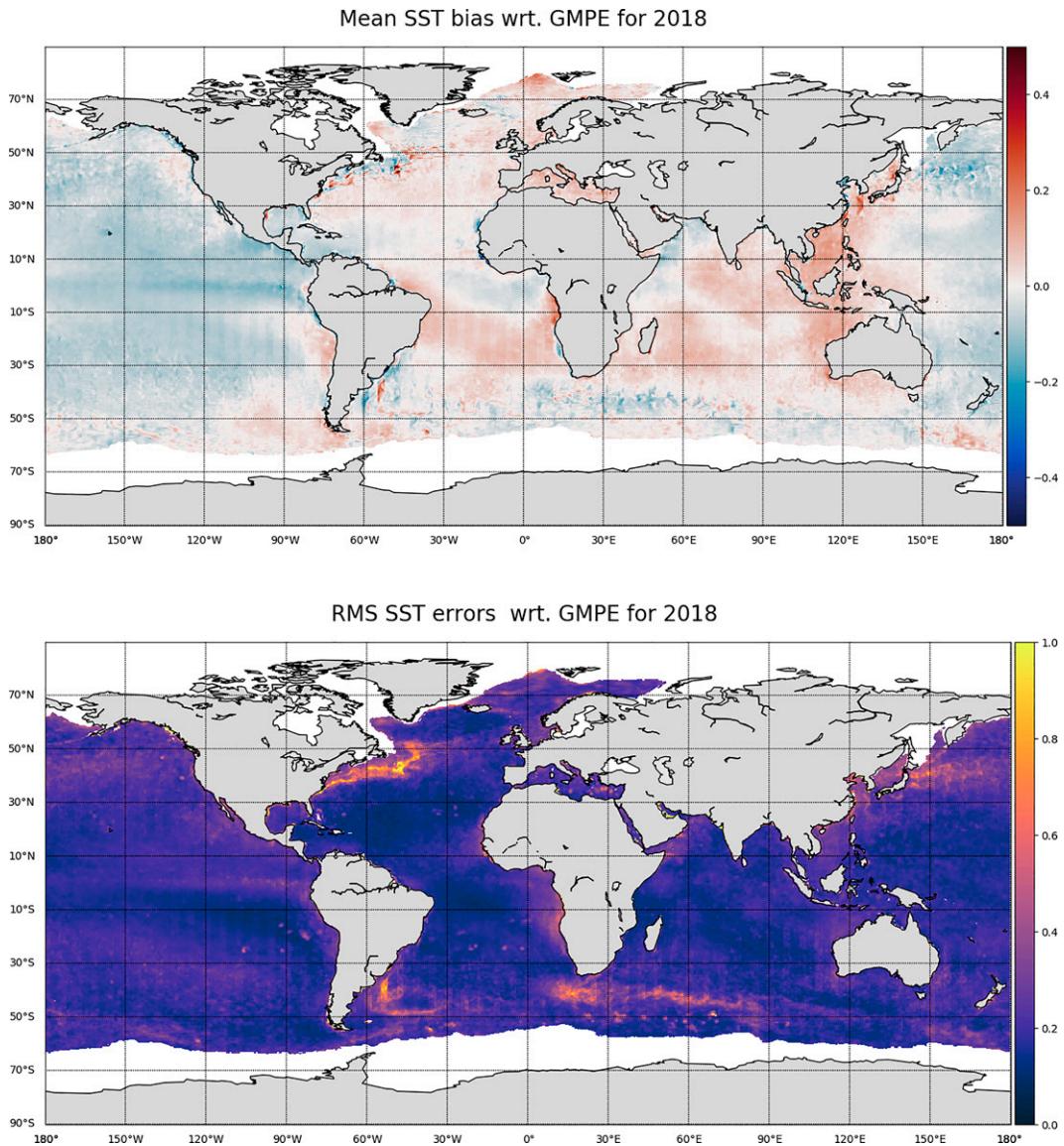


FIG. 6. Spatial distribution of the (top) annual mean errors and (bottom) RMS errors in SST with respect to the GMPE SST product.

whereas we have used a 7-day window. The good agreement between the estimates suggests that the assimilation procedure with the model forecast as a background field helps to spread information from sparse observations in both space and time efficiently. We point out that the SSH is not a prognostic field in HYCOM; rather, it is a diagnostic quantity that is obtained from the Montgomery potential. When adjusting the model with SLA data, the corrections are derived for layer thickness rather than SSH. The degree of agreement indicates that the strategy used here for assimilating SSH is effective.

Daily $1/4^\circ$ gridded maps of OISST are assimilated by the system. We compare the hindcast SST with the GHRST-GMPE median SST product. The spatial map of the mean error shows large regions of small bias in the model (Fig. 6,

top panel). Although this map indicates the presence of long-term differences over large regions, the magnitude is rather small, less than 0.1° on average, and rarely above 0.2° anywhere. The small mean bias with respect to assimilated observations is reassuring as no model-dependent heat flux corrections or relaxation was used. The shortwave and long-wave fluxes from ERA5 were adjusted using NASA CERES radiation flux estimates but these corrections by themselves are rarely adequate and can at times interfere adversely with the data assimilation process. Globally averaged RMS errors are less than 0.3°C for the assimilative run (Fig. 6, bottom panel). Overall, SST is well constrained, with a small bias and RMSD within specified observation errors. The exceptions are regions of high mesoscale activity such as the Gulf Stream

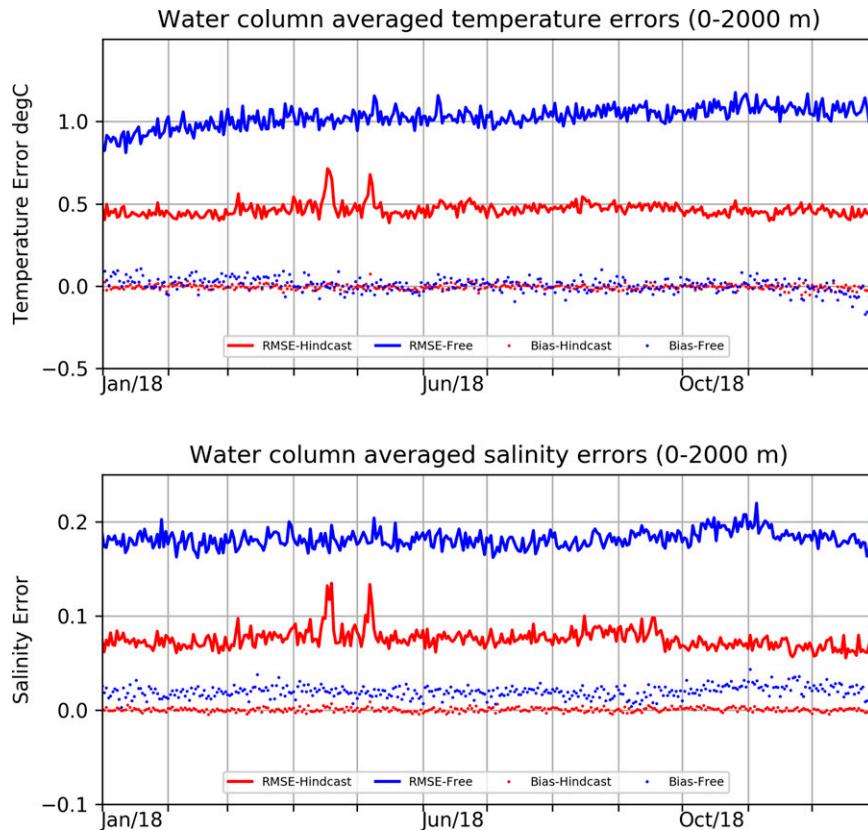


FIG. 7. Time series of globally averaged errors in (top) temperature and (bottom) salinity compared with assimilated Argo data. Model outputs and Argo profiles were interpolated to 32 standard levels (0–2000 m) for this comparison.

region where errors are on the order of 1.5°C but these errors are still well below the observed standard deviation of SST in these regions.

The assimilation of relatively coarsely sampled in situ T/S profiles is crucial for accurate representation of the vertical density structure of the ocean particularly when a large quantity of remotely sensed surface data are assimilated simultaneously. The hindcast assimilated T/S profiles from the Argo dataset and time series of globally averaged error compared with assimilated Argo T/S data show errors of magnitude 0.5°C for temperature and 0.1 for salinity in the upper 2000 m (Fig. 7). Between 0 and 500 m, departures from in situ observations are generally within 1.2°C and 0.2 for temperature and salinity, respectively. Exceptions occur in some high variability regions such as the Agulhas and Gulf Stream regions similar to situation with SLA and SST. Although these comparisons are not with entirely independent data, they are useful to gauge the consistency of the hindcast and the procedure used to assimilate profiles directly in generalized vertical coordinates. Recall that the T/S profiles were first converted from pressure coordinates to the generalized vertical coordinates before assimilation, a process that provides three quantities—layer temperature, layer salinity, and layer thickness—which are then assimilated layer by layer. The model outputs are then again interpolated from generalized coordinates to pressure

coordinates for these comparisons. The magnitude of the errors reported here are similar to ones reported in Oke et al. (2013), Lellouche et al. (2013), Blockley et al. (2014), Waters et al. (2015), and Martin et al. (2015) and therefore suggest that the approach adopted here for profile assimilation is effective.

Similar comparison with independent (withheld) T/S data from tropical moorings (Fig. 3) shows consistently lower errors for the hindcast compared to the free run (Fig. 8). These comparisons are restricted to the main thermocline region (0–500 m) where the largest errors generally occur. The magnitude of these errors for the hindcast are largely within 1.2°C and 0.2 for temperature and salinity, respectively, and are similar to the innovations reported in Lellouche et al. (2013) and Blockley et al. (2014). The errors are consistently below the errors for the free run suggesting a spread of information from observed to unobserved regions as expected for a data assimilative system. Space constraints do not permit an exhaustive analysis of the impact of T/S profiles here but we mention that statistics for the entire water column were computed in 2×2 bins and in vertical layers and show that the T/S profile assimilation reduces the large warm bias in the Indo-Pacific temperature and other temperature errors with complex spatial distribution. The fresh bias in salinity over most of the model domain is significantly reduced by the assimilation of T/S profiles.

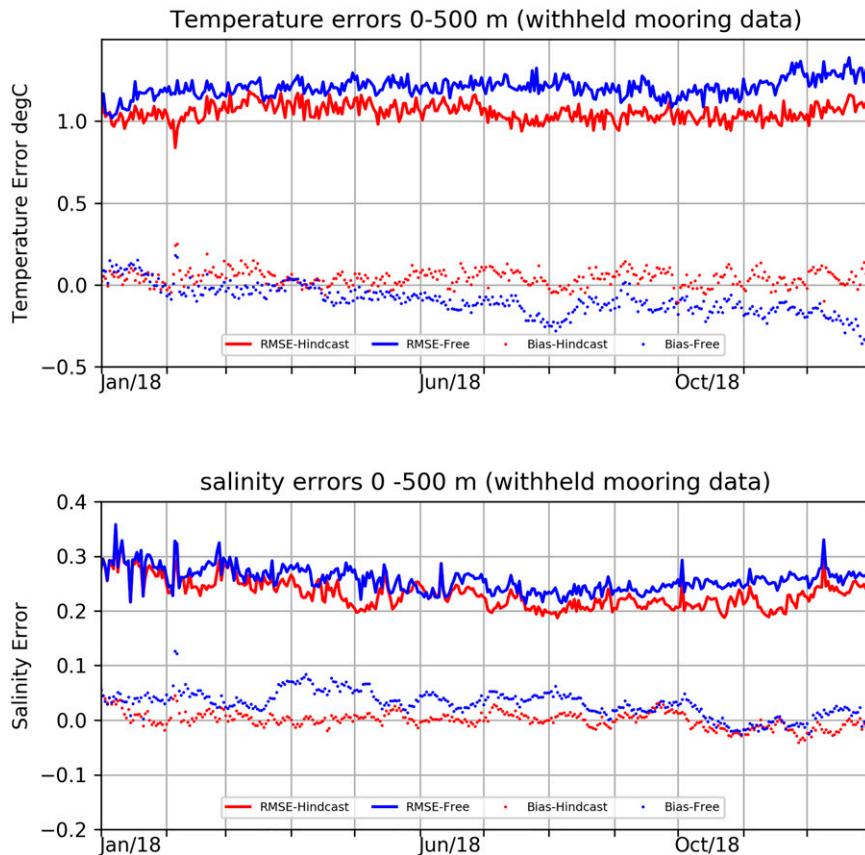


FIG. 8. Time series of globally averaged errors in (top) temperature and (bottom) salinity compared with withheld mooring data.

Finally, using interpolated velocities from the Global Drifter Program dataset, we compare the model results with the observed drifter velocities at 15-m depth and at 0000 UTC daily. These velocities were not assimilated and thus are a second set of independent observations for verification. Figure 9 shows a map of speed errors for the hindcast and the free run. The speed errors are reduced in all regions of the global ocean and particularly in regions of main ocean currents. Figure 10 shows a plot of the GAMMA parameter defined as ratio of the relative speed error. A lower GAMMA indicates better match between model and observations. The GAMMA values for the hindcast (around 0.7) are generally lower than the GAMMA values for the free running model (about 0.9). The RMS velocity errors (on average 20 cm s^{-1}) are lower than the observed standard deviation of the drifter velocities for the entire duration of the hindcast. Further, the correlations between drifter and model velocities are generally 15%–20% higher for the hindcast for the entire duration of the analysis and range between 0.5 and 0.6. We note that these correlations are statistically significant due to extremely large number of samples.

c. Hindcast summary

The results from the hindcast clearly show the positive impact of data assimilation in reducing errors compared to the

free run. The same initial conditions are used for both the free run and the hindcast. For the free run, errors increase slowly with time and saturate after about 3 months (Figs. 7, 8, 10). The hindcast is continuously constrained with data, and thus, the errors and correlations are more or less constant. Further, the magnitude of the errors are comparable to errors and innovations reported for similar systems (Oke et al. 2013; Lellouche et al. 2013; Blockley et al. 2014; Waters et al. 2015; Martin et al. 2015). Several aspects of the hindcast can be further improved. We have assimilated SLA data with a uniform error of 7 cm; this can be improved with spatially varying errors. Further, the vertical projection of SLA innovations is unlikely to be optimal everywhere, particularly in weakly stratified regions where alternate methods might be required. Similarly, the SST field was constrained with maps that provides information over the whole domain. A more robust test would be to assimilate the satellite and in situ directly. Finally, vertical profiles were converted from pressure coordinates to generalized vertical coordinates for assimilation. Altering both salinity and temperature is not straightforward in isopycnic part of the model and can lead to cabling. These aspects will have to be evaluated in longer-term hindcasts. The hindcast is being extended to three years (2018–20) and detailed verification will be reported elsewhere. Overall, the hindcast provides a good starting point for further improvements.

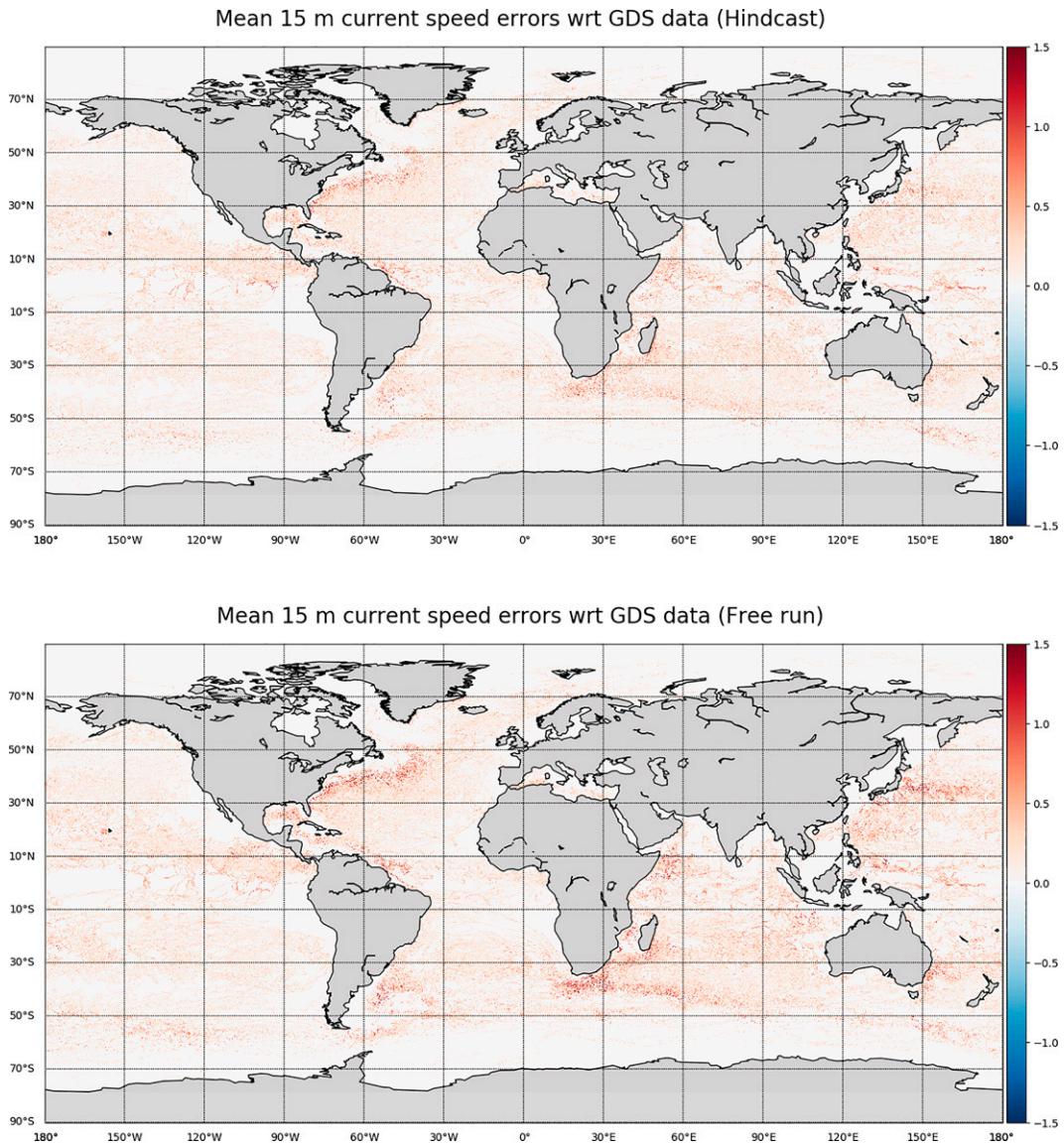


FIG. 9. Spatial distribution of errors in 15-m speed computed with respect to the GDP drifter dataset. (top) The error in the hindcast and (bottom) the free run.

7. Summary and conclusions

The purpose of this paper is to present an open source code for ocean data assimilation (T-SIS 2.0), provide some technical details of the implementation, and validate its performance in a global hindcast application. Specifying error covariances or the weighting matrices in the least squares formulation is the fundamental aspect of any data assimilation package. The package presented herein allows several ways to specify these matrices. Among them, the GMRF based method offers several advantages for large analysis problems. Analysis with millions of observations and model grid points can be computed with linear complexity in the number of grid points using iterative solvers without the need for partitioning the observations common to many other widely used schemes. Due to the ability of

relatively low-order GMRF to represent longer range correlations by successive local interactions, the GMRF scheme can be included as a subroutine within the ocean model without excessive modifications to the ocean model domain decomposition. In this case, data can be assimilated whenever available or at preset times without having to resort to input/output through restart files, which can have significant impact in long reanalysis type of computations.

The GMRF based information matrix \mathbf{L} used in this package are numerical discretizations of the diffusion operator. This can be related to the modeling of the inverse of the background error correlations by positive definite polynomials of the diffusion operator used in the variational assimilation methods (Yaremchuk and Smith 2011; Mirouze and Weaver 2010). The ability to specify inhomogeneity and anisotropy

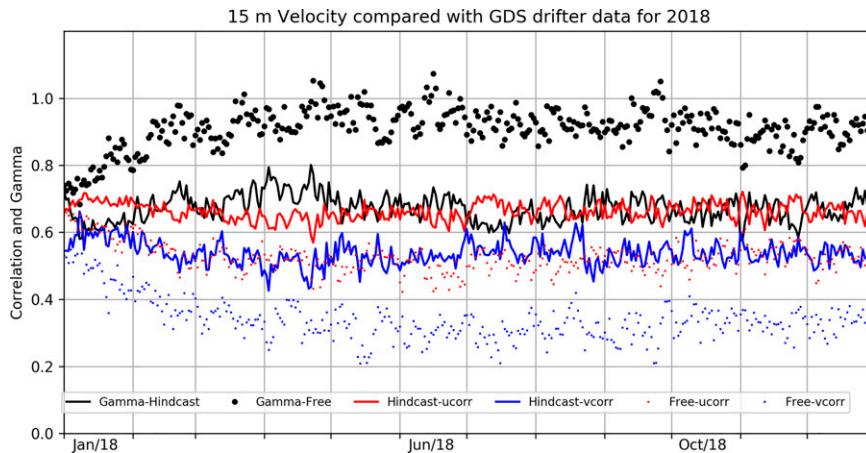


FIG. 10. Time series of GAMMA, defined as ratio of the relative speed error between observations and model, and correlation between modeled and observed drifter velocity components at 15 m.

via a diffusion tensor, the computational advantages, and the relatively easier enforcement of the positive definiteness are some of the advantages of variational approaches. These advantages are naturally present in the GMRF-based modeling of the implicit covariance.

The information matrices used here are limited to simple predefined adjacency matrices of the graph Laplacians. However, given the multiscale nature of the problem, a single GMRF parameterization is unlikely to work for all scales and undoubtedly other approaches to specify the “neighborhood” will be required, particularly to capture longer-range correlations. While correlation structures of a spatial random field and inverse factorizations of the covariance matrix is an active area of study (e.g., Yaremchuk et al. 2018; Yaremchuk and Sentchev 2012; Hsieh et al. 2014), our initial approach to model long range correlations is based on multiscale realizations. To capture a range of scales, a hierarchy of analysis grids can be set up going from coarse to fine resolutions with analysis starting from the coarse scales and going to finer scales and at each stage using the analysis from the previous resolution as the forecast/background values. This approach is akin to the classic successive corrections and is consistent with the standard modeling practice of forcing high-resolution models using the results of coarser-resolution models.

Apart from error covariance modeling, the performance of the data assimilative systems are crucially dependent on the observations, their preprocessing and quality control and many other technical details such as the vertical projection of the surface information, incremental updating, and in the case of HYCOM, handling additional complexity introduced by the hybrid nature of the vertical coordinate system. The package in its current version can assimilate all of the commonly available observations from remote sensing and in situ platforms such as SLA, SST, vertical profiles of temperature and salinity. The results from the hindcast using the package are similar to the performance reported for many operational

systems and provide a default starting point for choices that have to be made with respect to numerous implementation details that go into building a data assimilative system. Data assimilative systems based on this package are now in use for the Indian Ocean, South Atlantic, the Gulf of Mexico, and other areas and the package is continually updated based on the feedback provided by the user community. Recent updates include assimilation of velocity data from drifting buoys, HF radars, and assimilation of feature information such as locations of fronts estimated from ocean color. The latter is accomplished by a rubber sheeting process to modify the surface height field to align with the fronts which is then used to adjust the models density structure as with altimeter data.

Further work on the package is proceeding along three lines: 1) Multiscale nature of the analysis problem is being addressed by a multiresolution modeling approach wherein several fine-resolution models are embedded within a coarse global system. Each nested model has GMRF based error process that operates at a certain scale associated with the resolution. Interaction between models allow corrections at scales not well captured by the GMRF associated with a particular model; 2) a second aspect of the work relates to the recursion of the information matrix to evolve the error process in time. Chin et al. (1999) provides a starting point for the information matrix recursion which can then be updated periodically with information from ensembles to sidestep closure issues with tangent linear models. Finally, recent work in the statistics and machine learning community has begun to explore the connections between GMRF and convolutional neural networks (Siden and Lindsten 2020). In particular it has been pointed out that GMRFs similar to ones used in the package are essentially one-layer convolutional neural networks (CNN). This allows generalizing GMRFs to multilayer CNN architectures effectively increasing the scale of GMRF process with favorable computational scaling and possibly leading to deep GMRF based assimilation methodologies.

TABLE A1. Example water column projection of SLA increments.

Layer	Thickness before adjustment	Thickness after adjustment
1	3.001 384	3.001 384
2	3.541 663	3.541 663
3	4.179 188	4.179 188
4	10.45 275	127.0939
5	35.74 834	35.74 834
6	2.266 723	2.266 723
7	19.03 991	19.03 991
8	10.28 967	10.28 967
9	18.59 432	18.59 432
10	21.33 095	21.33 095
11	27.64 579	27.64 579
12	30.57 642	30.57 642
13	45.60 513	45.60 513
14	59.77 328	59.77 328
15	98.69 099	98.69 099
16	127.9392	127.9392
17	145.4958	145.4958
18	148.7269	148.7269
19	211.2244	211.2244
20	161.2507	161.2507
21	184.9853	184.9853
22	169.6162	169.6162
23	123.4931	123.4931
24	272.9174	272.9174
25	140.8281	140.8281
26	229.6139	229.6139
27	449.4691	449.4691
28	723.8014	723.8014
29	894.1266	894.1266
30	851.5503	734.9091
31	0.00×10^0	0.00×10^0
32	0.00×10^0	0.00×10^0

Acknowledgments. The present work was supported by ONR Grant N00014-19-1-2671. M. Iskandarani was also partially supported by NSF-1639722. Contributions by TMC was carried out at the Jet Propulsion Laboratory (JPL), California Institute of Technology, under a contract with the National Aeronautics and Space Administration (NASA). Discussions with Dr. W. C. Thacker are also gratefully acknowledged.

Data availability statement. Fortran programs for the test cases, data from the hindcast experiment, and the complete T-SIS package are available from <http://tendral.com/tsis/>.

APPENDIX

Vertical Projection of SLA Increments into the Water Column in T-SIS 2.0

The algorithm is a modification of the Cooper and Haines (1996) method for generalized vertical coordinates

used in HYCOM. It operates on a water column. Whenever possible, only two layers will have nonzero innovations, one above the main thermocline and one below. Their innovations are equal in magnitude and opposite in sign, so that the resulting pressure thickness for the water column will not change the bottom pressure. The magnitude of the innovations should guarantee that the sum of the corresponding pressure-thickness innovations is equal to the sea surface height innovation. If constraints on minimum allowed layer thickness prevents a sufficient part of one layer's pressure thickness to be transferred to the other to accommodate the SSH innovation, then that layer's innovation will be as large as allowed, and the procedure will be repeated for the adjacent layer. All layers beneath and including the layer containing the mixed layer base will be adjusted if necessary. To extend the adjustment into all fixed thickness layers at and beneath the mixed layer base, all layers between the layer containing the base and the first layer significantly thicker than the specified minimum thickness are consolidated into a single layer before the rearrangement. After the rearrangement, all layers within this consolidated layer are thinned or thickened in the same proportion as the consolidated layer. An example of the adjustment to a water column located at 12°S, 98°W for an SSH increment of 0.8 m is shown below in Table A1. Here layer 4 is inflated with water from layer 30 to match the observed sea level.

REFERENCES

- Anderson, J. L., 2012: Localization and sampling error correction in ensemble Kalman filter data assimilation. *Mon. Wea. Rev.*, **140**, 2359–2371, <https://doi.org/10.1175/MWR-D-11-00013.1>.
- , T. Hoar, K. Raeder, H. Liu, N. Collins, R. Torn, and A. Avellano, 2009: The Data Assimilation Research Testbed: A community facility. *Bull. Amer. Meteor. Soc.*, **90**, 1283–1296, <https://doi.org/10.1175/2009BAMS2618.1>.
- Besag, J., 1974: Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Stat. Soc.*, **36A**, 192–225, <https://doi.org/10.1111/j.2517-6161.1974.tb00999.x>.
- Bleck, R., 2002: An oceanic general circulation model framed in hybrid isopycnal-Cartesian coordinates. *Ocean Modell.*, **4**, 55–88, [https://doi.org/10.1016/S1463-5003\(01\)00012-9](https://doi.org/10.1016/S1463-5003(01)00012-9).
- Blockley, E. W., and Coauthors, 2014: Recent development of the Met Office operational ocean forecasting system: An overview and assessment of the new Global FOAM forecasts. *Geosci. Model Dev.*, **7**, 2613–2638, <https://doi.org/10.5194/gmd-7-2613-2014>.
- Bloom, S. C., L. L. Takachs, M. da Silva, and D. Ledvina, 1996: Data assimilation using incremental analysis updates. *J. Atmos. Oceanic Technol.*, **12A**, 1256–1271, [https://doi.org/10.1175/1520-0493\(1996\)124%3C1256:DAUIAU%3E2.0.CO;2](https://doi.org/10.1175/1520-0493(1996)124%3C1256:DAUIAU%3E2.0.CO;2).
- Brasseur, P., 2006: Ocean data assimilation using sequential methods based on the Kalman filter: From theory to practical implementations. *Ocean Weather Forecasting: An Integrated View of Oceanography*, E. Chassignet and J. Verron, Eds., Springer, 271–316.
- Chassignet, E. P., L. T. Smith, G. R. Halliwell Jr., and R. Bleck, 2003: North Atlantic simulations with the Hybrid Coordinate Ocean Model (HYCOM): Impact of vertical coordinate choice,

- reference pressure, and thermobaricity. *J. Phys. Oceanogr.*, **33**, 2504–2526, [https://doi.org/10.1175/1520-0485\(2003\)033<2504:NASWTH>2.0.CO;2](https://doi.org/10.1175/1520-0485(2003)033<2504:NASWTH>2.0.CO;2).
- , and Coauthors, 2009: U.S. GODAE: Global Ocean Prediction with the Hybrid Coordinate Ocean Model (HYCOM). *Oceanography*, **22** (2), 64–75, <https://doi.org/10.5670/oceanog.2009.39>.
- Chin, T. M., A. J. Mariano, and E. P. Chassignet, 1999: Spatial regression and multiscale approximations for sequential data assimilation in ocean models. *J. Geophys. Res.*, **104**, 7991–8014, <https://doi.org/10.1029/1998JC900075>.
- , A. J. Haza, and A. J. Mariano, 2002: A reduced-order information filter for multilayer shallow-water models: Profiling and assimilation of sea surface height. *J. Atmos. Oceanic Technol.*, **19**, 517–533, [https://doi.org/10.1175/1520-0426\(2002\)019<0517:AROFF>2.0.CO;2](https://doi.org/10.1175/1520-0426(2002)019<0517:AROFF>2.0.CO;2).
- Cooper, M., and K. Haines, 1996: Altimetric assimilation with water property conservation. *J. Geophys. Res.*, **101**, 1059–1077, <https://doi.org/10.1029/95JC02902>.
- Counillon, F., and L. Bertino, 2009: High-resolution ensemble forecasting for the Gulf of Mexico eddies and fronts. *Ocean Dyn.*, **59**, 83–95, <https://doi.org/10.1007/s10236-008-0167-0>.
- Crassidis, J. L., and J. L. Junkins, 2004: *Optimal Estimation of Dynamical Systems*. Chapman and Hall/CRC, 749 pp.
- Cummings, J. A., 2005: Operational multivariate ocean data assimilation. *Quart. J. Roy. Meteor. Soc.*, **131**, 3583–3604, <https://doi.org/10.1256/qj.05.105>.
- Daley, R., 1991: *Atmospheric Data Analysis*. Cambridge University Press, 457 pp.
- Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, **99**, 10143–10162, <https://doi.org/10.1029/94JC00572>.
- , 2003: The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dyn.*, **53**, 343–367, <https://doi.org/10.1007/s10236-003-0036-9>.
- , 2006: *Data Assimilation: The Ensemble Kalman Filter*. Springer, 280 pp.
- Fox-Kemper, B., and Coauthors, 2019: Challenges and prospects in ocean circulation models. *Front. Mar. Sci.*, **6**, 65, <https://doi.org/10.3389/fmars.2019.00065>.
- Halliwell, G. R., 2004: Evaluation of vertical coordinates and vertical mixing algorithms in the Hybrid-Coordinate Ocean Model (HYCOM). *Ocean Modell.*, **7**, 285–322, <https://doi.org/10.1016/j.ocemod.2003.10.002>.
- , A. Srinivasan, V. Kourafalou, H. Kang, D. Willey, M. Le Henaff, and R. Atlas, 2014: Rigorous evaluation of a fraternal twin ocean OSSE system in the open Gulf of Mexico. *J. Atmos. Oceanic Technol.*, **31**, 105–130, <https://doi.org/10.1175/JTECH-D-13-00011.1>.
- Heimbach, P., and Coauthors, 2019: Putting it all together: Adding value to the global ocean and climate observing systems with complete self-consistent ocean state and parameter estimates. *Front. Mar. Sci.*, **6**, 55, <https://doi.org/10.3389/fmars.2019.00055>.
- Hersbach, H., and Coauthors, 2018: ERA5 hourly data on single levels from 1979 to present. Copernicus Climate Change Service Climate Data Store, accessed 15 August 2018, <https://doi.org/10.24381/cds.adbb2d47>.
- Hollingsworth, A., and P. Lönnberg, 1986: The statistical structure of short range forecast errors as determined from rawinsonde data. Part I: The wind field. *Tellus*, **38A**, 111–136, <https://doi.org/10.1111/j.1600-0870.1986.tb00460.x>.
- Hsieh, C.-J., M. A. Sustik, I. S. Dhillon, and P. Ravikumar, 2014: Quadratic approximation for sparse inverse covariance estimation. *J. Mach. Learn. Res.*, **15**, 2911–2947, <https://jmlr.org/papers/v15/hsieh14a.html>.
- Jazwinski, A., 1970: *Stochastic Processes and Filtering Theory*. Academic Press, 376 pp.
- Kalnay, E., 2002: *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 341 pp., <https://doi.org/10.1017/CBO9780511802270>.
- Kato, S., and Coauthors, 2018: Surface irradiances of edition 4.0 Clouds and the Earth's Radiant Energy System (CERES) Energy Balanced and Filled (EBAF) data product. *J. Climate*, **31**, 4501–4527, <https://doi.org/10.1175/JCLI-D-17-0523.1>.
- Large, W. G., and S. G. Yeager, 2004: Diurnal to decadal global forcing for ocean and sea ice models: The data sets and flux climatologies. NCAR Tech. Note NCAR/TN-460+STR, 105 pp.
- , J. C. McWilliams, and S. C. Doney, 1994: Oceanic vertical mixing: A review and a model with a non-local boundary layer parameterization. *Rev. Geophys.*, **32**, 363–403, <https://doi.org/10.1029/94RG01872>.
- Lellouche, J. M., and Coauthors, 2013: Evaluation of global monitoring and forecasting systems at Mercator Ocean. *Ocean Sci.*, **9**, 57–81, <https://doi.org/10.5194/os-9-57-2013>.
- Le Traon, P., F. Nadal, and N. Ducet, 1998: An improved mapping method of multisatellite altimeter data. *J. Atmos. Oceanic Technol.*, **15**, 522–534, [https://doi.org/10.1175/1520-0426\(1998\)015<0522:AIMMOM>2.0.CO;2](https://doi.org/10.1175/1520-0426(1998)015<0522:AIMMOM>2.0.CO;2).
- Lindgren, F., H. Rue, and J. Lindström, 2011: An explicit link between Gaussian fields and Gaussian Markov random fields: The SPDE approach. *J. Roy. Stat. Soc.*, **73A**, 423–498, <https://doi.org/10.1111/j.1467-9868.2011.00777.x>.
- Lumpkin, R., and L. Centurioni, 2019: NOAA global drifter program quality-controlled 6-hour interpolated data from ocean surface drifting buoys. NOAA National Centers for Environmental Information, accessed 15 September 2021, <https://doi.org/10.25921/7ntx-z961>.
- Martin, M. J., and Coauthors, 2015: Status and future of data assimilation in operational oceanography. *J. Oper. Oceanogr.*, **8**, s28–s48, <https://doi.org/10.1080/1755876X.2015.1022055>.
- Mirouze, I., and A. T. Weaver, 2010: Representation of correlation functions in variational data assimilation using an implicit diffusion operator. *Quart. J. Roy. Meteor. Soc.*, **136**, 1421–1443, <https://doi.org/10.1002/qj.643>.
- Moore, A. M., and Coauthors, 2019: Synthesis of ocean observations using data assimilation for operational, real-time and reanalysis systems: A more complete picture of the state of the ocean. *Front. Mar. Sci.*, **6**, 90, <https://doi.org/10.3389/fmars.2019.00090>.
- Morrow, R., and Coauthors, 2019: Global observations of fine-scale ocean topography with the Surface Water and Ocean Topography (SWOT) mission. *Front. Mar. Sci.*, **6**, 232, <https://doi.org/10.3389/fmars.2019.00232>.
- Mulet, S., and Coauthors, 2021: The new CNES-CLS18 global mean dynamic topography. *Ocean Sci.*, **17**, 789–808, <https://doi.org/10.5194/os-17-789-2021>.
- Oke, P. R., J. S. Allen, R. N. Miller, G. D. Egbert, and P. M. Kosro, 2002: Assimilation of surface velocity data into a primitive equation coastal ocean model. *J. Geophys. Res.*, **107**, 3122, <https://doi.org/10.1029/2000JC000511>.

- , and Coauthors, 2013: Towards a dynamically balanced eddy-resolving ocean reanalysis: BRAN3. *Ocean Modell.*, **67**, 52–70, <https://doi.org/10.1016/j.ocemod.2013.03.008>.
- Paciorek, C. J., 2013: Spatial models for point and areal data using Markov random fields on a fine grid. *Electron. J. Stat.*, **7**, 946–972, <https://doi.org/10.1214/13-EJS791>.
- Panteleev, G., M. Yaremchuk, J. Stroh, P. Posey, D. Herbert, and D. Nechaev, 2015: Optimization of high-frequency radar sites in the Bering Strait region. *J. Atmos. Oceanic Technol.*, **32**, 297–309, <https://doi.org/10.1175/JTECH-D-14-00071.1>.
- Piolle, J. F., E. Autret, O. Arino, I. S. Robinson, and P. Le Borgne, 2010: Medspiration, toward the sustained delivery of satellite SST products and services over regional seas. *Proc. 2010 ESA Living Planet Symp.*, Bergen, Norway, ESA.
- Rue, H., and H. Tjelmeland 2002: Fitting Gauss Markov random fields to Gaussian fields. *Scand. J. Stat.*, **29**, 31–49, <https://doi.org/10.1111/1467-9469.00058>.
- , and L. Held, 2005: *Gaussian Markovian Random Fields: Theory and Applications*. Chapman and Hall/CRC, 280 pp.
- Siden, P., and F. Lindsten, 2020: Deep Gaussian Markov random fields. arXiv, <https://arxiv.org/abs/2002.07467>.
- Srinivasan, A., and Coauthors, 2011: A comparison of sequential assimilation schemes for ocean prediction with the Hybrid Coordinate Ocean Model (HYCOM): Twin experiments with static forecast error covariance. *Ocean Modell.*, **37**, 85–111, <https://doi.org/10.1016/j.ocemod.2011.01.006>.
- Thacker, W. C., 2008: Estimating salinity between 25° and 45°S in the Atlantic Ocean using local regression. *J. Atmos. Oceanic Technol.*, **25**, 114–130, <https://doi.org/10.1175/2007JTECH0530.1>.
- Troupin, C., and Coauthors, 2012: Generation of analysis and consistent error fields using the Data Interpolating Variational Analysis (DIVA). *Ocean Modell.*, **52–53**, 90–101, <https://doi.org/10.1016/j.ocemod.2012.05.002>.
- Waters, J., D. J. Lea, M. J. Martin, D. Storkey, and J. While, 2015: Implementing a variational data assimilation system in an operational 1/4 degree global ocean model. *Quart. J. Roy. Meteor. Soc.*, **141**, 333–349, <https://doi.org/10.1002/qj.2388>.
- Wunsch, C., 1996: *The Ocean Circulation Inverse Problem*. Cambridge University Press, 442 pp.
- Xu, F., and A. Ignatov, 2014: In situ SST Quality Monitor (iQuam). *J. Atmos. Oceanic Technol.*, **31**, 164–180, <https://doi.org/10.1175/JTECH-D-13-00121.1>.
- Yaremchuk, M., and S. Smith, 2011: On the correlation functions associated with the polynomials of the diffusion operator. *Quart. J. Roy. Meteor. Soc.*, **137**, 1927–1932, <https://doi.org/10.1002/qj.893>.
- , and G. Sentchev, 2012: Multi-scale correlation functions associated with polynomials of the diffusion operator. *Quart. J. Roy. Meteor. Soc.*, **138**, 1948–1953, <https://doi.org/10.1002/qj.1896>.
- , J. D’Addezio, G. Panteleev, and G. Jacobs, 2018: On the approximation of the inverse error covariance of high resolution satellite altimetry data. *Quart. J. Roy. Meteor. Soc.*, **144**, 1995–2000, <https://doi.org/10.1002/qj.3336>.