An adaptive optimal interpolation based on analog forecasting: application to SSH in the Gulf of Mexico

Yicun Zhen, Pierre Tandeo, Stéphanie Leroux, Sammy Metref, Thierry Penduff, Julien Le Sommer

An adaptive optimal interpolation based on analog forecasting: application to SSH in the Gulf of Mexico

Yicun Zhen* and Pierre Tandeo

IMT Atlantique, Lab-STICC, UBL, Brest, France

Stéphanie Leroux

Ocean-Next, Grenoble, France

Sammy Metref, Thierry Penduff and Julien Le Sommer

Université Grenoble Alpes, CNRS, IRD, IGE, Grenoble, France.

*Corresponding author address: Dept. Signal & Communications, IMT Atlantique, 655 Avenue du Technopole, 29200 Plouzané, France

E-mail: zhenyicun@protonmail.com

Early Online Release: This preliminary version has been accepted for publication in *Journal of Atmospheric and Oceanic Technology*, may be fully cited, and has been assigned DOI 10.1175/JTECH-D-20-0001.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

ABSTRACT

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

Because of the irregular sampling pattern of raw altimeter data, many oceanographic applications rely on information from sea surface height (SSH) products gridded on regular grids where gaps have been filled with interpolation. Today, the operational SSH products are created using the simple, but robust, optimal interpolation (OI) method. If well tuned, the OI becomes computationally cheap and provides accurate results at low resolution. However, OI is not adapted to produce high resolution and high frequency maps of SSH. To improve the interpolation of SSH satellite observations, a data-driven approach (i.e., constructing a dynamical forecast model from the data) was recently proposed: analog data assimilation (AnDA). AnDA adaptively chooses analog situations from a catalog of SSH scenes – originating from numerical simulations or a large database of observations – which allow the temporal propagation of physical features at different scales, while each observation is assimilated. In this article, we review the AnDA and OI algorithms and compare their skills in numerical experiments. The experiments are observing system simulation experiments (OSSE) on the Lorenz-63 system and on an SSH reconstruction problem in the Gulf of Mexico. The results show that AnDA, with no necessary tuning, produces comparable reconstructions as does OI with tuned parameters. Moreover, AnDA manages to reconstruct the signals at higher frequencies than OI. Finally, an important additional feature for any interpolation method is to be able to assess the quality of its reconstruction. This study shows that the standard deviation estimated by AnDA is flow-dependent, hence more informative on the reconstruction quality, than the one estimated by OI.

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

1. Introduction

Satellite altimetry is an essential component of the global ocean observing system with many applications key to climate monitoring, operations at sea and oceanic process understanding. Satellite altimeters provide measurements of sea surface height (SSH), a dynamical parameter that holds information about the upper ocean pressure field. Satellite derived SSH measurements are used for monitoring changes in sea-level at global and regional scales. They are also used for estimating upper ocean circulation at scales larger than the first Rossby radius of deformation where the geostrophic balance holds. Satellite altimetry is therefore a key source of information for ocean monitoring systems, and an essential constraint in ocean forecasting systems.

In practice, many oceanographic applications of satellite altimetry rely on gridded SSH products rather than on raw along-track SSH data. Satellite altimeters indeed provide SSH measurements along ground tracks, following a sampling pattern which depend on the satellite orbit. The existing constellation of altimeters combines several satellites, but the overall sampling of SSH data is irregular with large gaps both in space and in time, and will remain so in the near future with the advent of wide-swath altimetry. Still, many applications of SSH data require the tracking of oceanic flow features in space and time or the computation of spatial derivatives of SSH such as applications related to ship routing, search and rescue, oil spills, or fisheries, as detailed in Le Traon et al. (2019). Hence, for convenience, many applications of SSH data are currently based on operational data products where SSH data is interpolated on a regular spatial grid at fixed time intervals.

Presently, the most commonly used operational gridded SSH products are based on static interpolation methods. Operational gridded SSH L4 products, as distributed for instance by the AVISO data center within the Copernicus programme (Pujol et al. 2016; Le Traon et al. 2019), combine

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

information from multiple altimeters through an optimal interpolation (OI) analysis. Optimal interpolation analysis (Gandin 1965) is a static interpolation method which uses the autocorrelation of a field to define the relative weights given to a set of observed data for reconstructing the field at unobserved locations. In practice, gridded SSH products are therefore obtained as weighted sums of observed SSH values, derived from explicit assumptions as to the space and time autocorrelation structure of the SSH field.

Although widely used, OI-based gridded SSH products are affected by several limitations and shortcomings. The quality of OI-based SSH reconstructions is indeed intrinsically dependent on the choice of the predefined autocorrelation parameters; but in practice, the chosen autocorrelation parameters are usually not optimal because of the tradeoffs due to the optimization of the product resolution at global scale (Dibarboure et al. 2011; Pujol et al. 2012). Moreover, the OI procedure does not provide an a priori estimation of the level of error of the reconstructed fields. Most importantly, OI is not state dependent and therefore does not account for the complex, non-linear dynamics of oceanic flows (Ubelmann et al. 2015). These limitations and shortcomings will likely become more problematic with the higher spatial resolution capability of upcoming wide-swath altimeters (Fu and Ferrari 2008; Durand et al. 2010).

Several alternative approaches to static interpolation methods have been proposed in the context of ocean remote sensing. Methods have for instance been proposed for improving the representation, and estimation of the covariance structure of the field to interpolate. This includes the DINEOF method (Beckers and Rixen 2003), a parameter-free procedure used for interpolating sea surface temperature (SST) or surface chlorophyll (Chl). In the context of SSH mapping, approaches accounting explicitly for the nonlinear dynamics of SSH have been proposed. Ubelmann et al. (2015) relies for instance on a dynamical propagator based on quasi-geostrophic theory. Alternatively, Lguensat et al. (2019) proposes to use analog forecasting for accounting for ocean

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

dynamics in SSH mapping algorithms. Research have also focused on exploiting synergies between different sensors for improving SSH mapping algorithms (as for instance with SST, see Fablet et al. 2018a).

Because it is parameter-free and state dependent, Analog Data Assimilation appears as a promising approach for improving SSH mapping algorithms. Analog Data Assimilation (AnDA), also known as empirical dynamical modelling, is a state estimation procedure which combines data assimilation and analog forecasting (Tandeo et al. 2015; Lguensat et al. 2017). AnDA uses a catalog of trajectories in the system state space, which can be drawn from observations or from numerical model simulations. The catalog is used for inferring the system dynamics and for building estimates of the system state at unobserved locations and times. Realistic applications to oceanic data include the interpolation of SST (Fablet et al. 2018b) and the interpolation of SSH (Lguensat et al. 2019). Lguensat et al. (2019) have shown in particular how AnDA can be used for improving OI-based SSH fields at fine scale. Still, to date, a comparison of the respective skills and performances of OI versus AnDA in the context of SSH mapping is still missing.

In this study, we investigate how AnDA performs as compared to OI for the reconstruction of SSH maps from along-track SSH data. Our aim is to document the potential benefits of AnDA in the context of the design of operational gridded L4 SSH products. We present results based on Observing System Simulation Experiments (OSSE) over the Gulf of Mexico where the true state and the catalog of scenes are drawn from different members of a 50 members, ensemble model simulation run at 1/4° resolution. Our analysis focuses in particular on the relative performance of AnDA and OI in reconstructing the time variability of SSH signals, on the sensitivity of the reconstruction to the size of the catalog and the ability of the methods to estimate the quality of their recontructions.

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

Within the limitations of our OSSE experiments, our results show that : (i) AnDA provides estimates of SSH with error levels comparable to an optimally tuned OI but without the need to a priori tune the covariance parameters; (ii) AnDA can reconstruct more reliably high frequency SSH fluctuations than OI, which shows limited skill for time-scales faster than the pre-tuned temporal correlation (iii) AnDA provides a reliable a priori estimate of the absolute error of the reconstructed SSH field, therefore allowing to detect when the quality of the reconstruction is poor. Our results therefore suggest that applications of AnDA to the mapping of SSH are worth investigating further.

This paper is organized as follows. In section 2, OI and AnDA algorithms are respectively reviewed, and details are given on how to tune the parameters. Then, both methods are applied to the Lorenz-63 system in section 3. Finally, in section 4, AnDA and OI are implemented on the SSH mapping problem in the region of the Gulf of Mexico. Section 5 brings a summary and a final discussion and conclusions. The code and data for reproducing the numerical results of the SSH experiments are available online (Zhen et al. 2019).

2. Description of the interpolation algorithms

OI is a widely used method for interpolating sparse and noisy observations. On the other hand, a data-driven interpolation method (i.e., constructing a dynamical forecast model from the data) AnDA has been introduced by Tandeo et al. (2015) and described in details by Lguensat et al. (2017). The details of these two algorithms are the following.

a. Optimal interpolation

OI is written as a linear inverse problem such as

$$\begin{cases} \mathbf{x} = \mathbf{x}^b + \boldsymbol{\eta}^b & (1) \\ \mathbf{y} = \mathbf{H}\mathbf{x} + \boldsymbol{\epsilon} & (2) \end{cases}$$

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

with $\mathbf{x}^b$ the background or a priori information, $\mathbf{H}$ the transformation from state $\mathbf{x}$ to observations $\mathbf{y}$, $\boldsymbol{\eta}^b \sim \mathcal{N}(\mathbf{0}, \mathbf{B})$ the background error and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$ the observation error. Here, $\mathbf{x}^b$, $\mathbf{B}$ and $\mathbf{R}$ are prescribed by the users. OI is a reanalysis and has a direct Gaussian solution given by $\mathcal{N}(\mathbf{x}^s, \mathbf{P}^s)$ such that

$$\begin{aligned} \mathbf{x}^s &= \mathbf{x}^b + \mathbf{K}\left(\mathbf{y} - \mathbf{H}\mathbf{x}^b\right) \\ \mathbf{P}^s &= \mathbf{B} - \mathbf{K}\mathbf{H}\mathbf{B} \end{aligned} \tag{3}$$

with $\mathbf{K} = \mathbf{B}\mathbf{H}^\top \left(\mathbf{H}\mathbf{B}\mathbf{H}^\top + \mathbf{R}\right)^{-1}$ the gain controlling the influence of the observations and the background.

The quality of OI results largely depends on the choice of the $\mathbf{B}$ and $\mathbf{R}$ matrices (Tandeo et al. 2018). The matrix $\mathbf{R}$ represents the error covariances in the observational model. It can be measured or estimated off-line if we assume that the observation error is stationary, which is the case in this article. However, in realistic applications, $\mathbf{R}$ can be non-stationary and should be estimated online (Minamide and Zhang 2017). The matrix $\mathbf{R}$ is not necessarily diagonal, i.e., the observation errors can be correlated. But, in practice, $\mathbf{R}$ is often assumed diagonal in order to reduce computational costs (Miyoshi et al. 2013). In our experiment, we set

$$\mathbf{R} = r\mathbf{I}, \tag{4}$$

where r is a scalar and $\mathbf{I}$ is the identity matrix.

The choice of $\mathbf{B}$ should be consistent with the choice of $\mathbf{x}^b$. If $\mathbf{x}^b$ is chosen to be the climatological mean state field $\bar{\mathbf{x}}$, then it is reasonable to choose $\mathbf{B}$ as the spatial-temporal climatology background covariance matrix. However, saving the complete spatial-temporal climatology covariances is not possible in large dimensional applications because of the prohibitive requirement for storage space. Therefore, a parameterized covariance matrix is often used to substitute the complete climatology covariances (Wu 1995; Gaspari and Cohn 1999). A popular choice of $\mathbf{B}$ has

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

the following form:

$$\mathbf{B}(x_{i,t_1}, x_{j,t_2}) = \mathbf{B}^{\text{spatial}}(i,j) f(d_t/L_t), \tag{5}$$

with $dt = |t_1 - t_2|$ and where $\mathbf{B}^{\text{spatial}}(i,j)$ is the (i,j)-th component of a pre-determined symmetric positive-definite matrix that represents the spatial climatology distribution of the state variable $\mathbf{x}$, f is a pre-determined function that defines the shape of the temporal correlation of each component of $\mathbf{x}$ and L_t is a prescribed parameter that defines a uniform decay rate for the temporal correlation. The matrix $\mathbf{B}^{\text{spatial}}$ can be a parametrized matrix or the sample covariances computed from a long time series of $\mathbf{x}$. Technically, $\mathbf{B}$ must be a symmetric positive-definite matrix. Hence, the choice of f can not be arbitrary. When the dimension of $\mathbf{x}$ is large, directly inverting the full matrix $\mathbf{HBH}^\top + \mathbf{R}$ is numerically demanding. In the present study, we implement OI locally in the spatial dimension, as presented in Algorithm 1. The choice of $\mathbf{B}^{\text{spatial}}$ and f depends on the application problem and will be discussed in each experimental section. Note that OI can also be implemented locally in both spatial and temporal dimensions.

b. Analog data assimilation

AnDA is a combination of analog forecasting and data assimilation. For the part of data assimilation, we use the ensemble Kalman smoother (EnKS), which is commonly used in many classic data assimilation problems (see for instance Compo et al. 2011). The EnKS requires an ensemble run of N_e simulations starting from different initial states. This ensemble run provides sample covariances for data assimilation at every time step. The EnKS consists of a forward filter and a backward smoother. In the forward process, the forecast of each ensemble member is calculated separately. And each member is updated by ensemble Kalman filter whenever observations are available. In the backward smoother, each member is updated recursively in the backward direction. The EnKS is summarized in Algorithm 2. The subscript i refers to the i-th member, t the

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

Algorithm 1 Local Optimal Interpolation

$t = 1, ..., T$, n_x is the spatial dimension of $\mathbf{x}$.

Input:$\bar{\mathbf{x}}$,$\mathbf{B}$,$\mathbf{H}$,$\mathbf{y}$, r, L_t, L_x

Output:$\hat{\mathbf{x}}_t^s$ and $\mathbf{P}_t^s$(only the diagonal elements)

1: **for** i_x=1,2,...,n_x **do**:

2: Let $\mathscr{N}^b$ be the collection of grid points whose distance from the i_x-th grid point is less than L_x.

3: **Get $\mathbf{x}_{loc}^b$:** restrict our attention to $\mathscr{N}^b$. Construct $\mathbf{x}_{loc}^b$ based on the climatology mean $\bar{\mathbf{x}}$ and $\mathscr{N}^b$, which represents the background estimate at all the grid points inside $\mathscr{N}^b$, from $t = 1$ to $t = T$.

4: **Get $\mathbf{B}_{loc}$:** construct $\mathbf{B}_{loc}$ which is the restriction of $\mathbf{B}$ to the state variables inside $\mathscr{N}^b$.

5: **Get $\mathbf{y}_{loc}$ and $\mathbf{R}_{loc}$:** construct the corresponding $\mathbf{y}_{loc}$, which consists of all the observations located inside $\mathscr{N}^b$ from $t = 1$ to $t = T$. Construct the corresponding $\mathbf{R}_{loc} = r\mathbf{I}_{loc}$, where the dimension of $\mathbf{I}_{loc}$ equals the dimension of $\mathbf{y}_{loc}$.

6: **Get $\mathbf{H}_{loc}$:** construct the corresponding $\mathbf{H}_{loc}$, which maps $\mathbf{x}_{loc}^b$ to the space of $\mathbf{y}_{loc}$

7: Calculate $\hat{\mathbf{x}}_{loc}^s$ and $\mathbf{P}_{loc}^s$ based on $\mathbf{x}_{loc}^b, \mathbf{y}_{loc}, \mathbf{H}_{loc}, \mathbf{B}_{loc}$, and Eq. (3).

8: Assign the value of $\hat{\mathbf{x}}_{loc}^s$ at the i_x-th grid point to the i_x-th component of $\hat{\mathbf{x}}^s$.

9: Assign the i_x-th diagonal component of $\hat{P}^s$ the variance of state variable at the i_x-th grid point inferred by $\hat{P}_{loc}^s$.

time, and the superscripts p, f, a, s refer to the foreacast without noise, forecast with noise, analysis, and reanalysis, respectively. In the forward Kalman filter, $\varepsilon_{i,t}$ is artificially created and added to $\mathbf{y}_t$ to compensate for the loss of variance (Houtekamer and Mitchell 1998). Line 4 of Algorithm 2 implements covariance localization which consist in the Schur product $\mathbf{P} \circ \mathbf{C}_{loc}$ with $\mathbf{C}_{loc}$ a

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

prescribed spatial covariance localization matrix (e.g., the Gaussian function or the Gaspari-Cohn matrix, Gaspari and Cohn 1999). In AnDA, the forecasting operator F in line 7 of Algorithm 2 is replaced by the analog-forecasting.

The major difference between AnDA and the classic data assimilation is that AnDA uses the technique of analog forecasting to predict the state at the next time step, instead of running the numerical model. In many applications, the analog forecast method could be an interesting alternative since it can simulate variable dynamics that are not necessarily represented in a numerical model. For instance, if an underlying variable of the system is not modeled by the numerical model but is present in the analog database, the analog forecast will be able to describe its relationship to other variables and predict its evolution. To insure the good performances of the analog forecast method and consequently of AnDA, a large historical dataset of state variables is needed: the catalog.

The quality of the analog forecasting procedure highly depends on the quality and the space of the catalog. Firstly, the catalog has to be as rich as possible to cover all the possible situations. Larger catalogs usually lead to better performance of AnDA. Secondly, the analogs have to live in an informative space. In practice, it can be a subspace to reduce the dimensionality of the problem (e.g., the EOF space used in section 4) or an augmented space when the dimension of the system is too low to distinguish situations that are not real analogs (e.g., the time delayed state space used in section 3). The catalog is then saved in a k-dimensional tree structure so that the relevant analogs at each time step can be accessed efficiently (Bentley 1975). The technique of analog forecasting at each time step can be briefly summarized by the following three steps.

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

Algorithm 2 Ensemble Kalman Smoother (with Covariance Localization)

$t = 1, ..., T$ and $i = 1, ..., N_e$.

Input:$\mathbf{x}^f_{i,1}$, $\mathbf{H}_t, \mathbf{R}, F(\cdot)$, $\mathbf{y}_t$,$\mathbf{C}_{loc}$

Output:$\hat{\mathbf{x}}^s_t$, $\mathbf{P}^s_t$

The forward ensemble Kalman filter

for $t = 1, 2, ..., T$ **do**:

$\bar{\mathbf{x}}^f_t \leftarrow \frac{1}{N_e} \sum_{i=1}^{N_e} \mathbf{x}^f_{i,t}$

$\mathbf{P}^f_t \leftarrow \frac{1}{N_e - 1} \sum_{i=1}^{N_e} (\mathbf{x}^f_{i,t} - \bar{\mathbf{x}}^f_t)(\mathbf{x}^f_{i,t} - \bar{\mathbf{x}}^f_t)^\top$

$\mathbf{P}^f_t \leftarrow \mathbf{P}^f_t \circ \mathbf{C}_{loc}$ (covariance localization)

$\mathbf{K}_t \leftarrow \mathbf{P}^f_t \mathbf{H}_t^\top (\mathbf{R} + \mathbf{H}_t \mathbf{P}^f_t \mathbf{H}_t^\top)^{-1}$

Draw $\varepsilon_{i,t} \sim \mathcal{N}(0, \mathbf{R})$

$\mathbf{x}^a_{i,t} \leftarrow \mathbf{x}^f_{i,t} + \mathbf{K}_t (\mathbf{y}_t - \mathbf{H}_t \mathbf{x}^f_{i,t} + \varepsilon_{i,t})$

$\mathbf{x}^f_{i,t+1}, \mathbf{x}^p_{i,t+1} \leftarrow F(\mathbf{x}^a_{i,t})$, forecast the state at $t+1$. $\mathbf{x}^p_{i,t+1}$ is the forecast without adding noises. When EnKS is applied within AnDA, F is replaced by the analog forecast.

The backward ensemble Kalman smoother

$\mathbf{x}^s_{i,T} \leftarrow \mathbf{x}^a_{i,T}$

for t=T-1,T-2,...,1 **do**:

$\bar{\mathbf{x}}^a_t \leftarrow \frac{1}{N_e} \sum_{i=1}^{N_e} \mathbf{x}^a_{i,t}$

$\bar{\mathbf{x}}^p_{t+1} \leftarrow \frac{1}{N_e} \sum_{i=1}^{N_e} \mathbf{x}^p_{i,t+1}$

$\mathbf{S}^a_t = (\mathbf{x}^a_{1,t} - \bar{\mathbf{x}}^a, ..., \mathbf{x}^a_{N_e,t} - \bar{\mathbf{x}}^a)$

$\mathbf{S}^p_{t+1} = (\mathbf{x}^p_{1,t+1} - \bar{\mathbf{x}}^p, ..., \mathbf{x}^p_{N_e,t+1} - \bar{\mathbf{x}}^p)$

$\mathbf{P}^a_t \leftarrow \mathbf{S}^a_t (\mathbf{S}^a_t)^\top / (N_e - 1)$

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

16: $\mathbf{P}_t^a \leftarrow \mathbf{P}_t^a \circ \mathbf{C}_{loc}$ (covariance localization)

17: $\mathbf{F}_{t+1} \leftarrow \mathbf{S}_{t+1}^p (\mathbf{S}_t^a)^\dagger$

18: $\mathbf{A}_t \leftarrow \mathbf{P}_t^a (\mathbf{F}_{t+1})^\top$

19: $\mathbf{J}_t \leftarrow \mathbf{A}_t (\mathbf{P}_t^f)^{-1}$

20: $\mathbf{x}_{i,t}^s \leftarrow \mathbf{x}_{i,t}^a + \mathbf{J}_t(\mathbf{x}_{i,t+1}^s - \mathbf{x}_{i,t+1}^f)$

21: $\hat{\mathbf{x}}_t^s \leftarrow \frac{1}{N_e} \sum_{i=1}^{N_e} \mathbf{x}_{i,t}^s$

22: $\mathbf{P}_t^s \leftarrow \frac{1}{N_e - 1} \sum_{i=1}^{N_e} (\mathbf{x}_{i,t}^s - \bar{\mathbf{x}}_t^s)(\mathbf{x}_{i,t}^s - \bar{\mathbf{x}}_t^s)^\top$

- Step 1: for a given state estimate $\mathbf{x}_t$, search for k analogs $(\mathbf{A}_1, ..., \mathbf{A}_k)$ that are nearest to $\mathbf{x}_t$ within the catalog, where k is pre-chosen. At the same time, we are also given the successors of $\mathbf{A}_i$'s, denoted by $\mathbf{S}_1, ..., \mathbf{S}_k$. Here $\mathbf{S}_i$ is the physical state at one time step later than $\mathbf{A}_i$.

- Step 2: build a local model $\mathscr{M}_t$ between $\mathbf{A}_1, ..., \mathbf{A}_k$ and $\mathbf{S}_1, ..., \mathbf{S}_k$, i.e. $\mathbf{S}_i = \mathscr{M}_t(\mathbf{A}_i) + \eta_{i,t}$, where $\eta_{i,t}$ is assumed to be some white and independent identically distributed noise, the distribution of which can be calculated from $\mathbf{A}_i$'s and $\mathbf{S}_i$'s.

- Step 3: apply the local model $\mathscr{M}_t$ to $\mathbf{x}_t$: $\mathbf{x}_{t+1} \leftarrow \mathscr{M}_t(\mathbf{x}_t) + \eta_t$, where η_t, describing the model error of $\mathscr{M}_t$, is drawn randomly and follows the same distribution as $\eta_{i,t}$.

It has been pointed out in Lguensat et al. (2017) that there are various choices of local models in the second step. Lguensat et al. (2019) compared these local models and the numerical results show that the locally linear model outperforms the others. In our applications, the local model $\mathscr{M}_t$ is the locally linear model that regresses $\mathbf{S}_i$ over the anomalies of analogs $\mathbf{A}_i' = \mathbf{A}_i - \bar{\mathbf{A}}$, where $\bar{\mathbf{A}}$ refers to the weighted mean of $\mathbf{A}_i$'s. Or equivalently, the local model we choose is the linear model that regresses the anomalies of successors $\mathbf{S}_i' = \mathbf{S}_i - \bar{\mathbf{S}}$ over $\mathbf{A}_i'$. In the numerical implementation, this linear regression can be done with respect to the leading components of $\mathbf{A}_i'$. In the case that

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

$\mathbf{x}_t$ represents the full state, this local model can be thought of as an approximation of the tangent linear model restricted on the attractor if the current state estimate $\mathbf{x}_t$ lies on the attractor and the distribution of analogs is dense enough. Furthermore, the distribution of the residuals $\eta_{i,t}$ is always assumed to be Gaussian in our applications. Hence, $\eta_t \sim \mathcal{N}(0, \mathbf{Q}_t)$, where $\mathbf{Q}_t$ is the weighted covariance matrix of the residues $\mathbf{S}_i - \mathscr{M}_t(\mathbf{A}_i)$, mentioned in step 3. The details of analog forecast with locally linear model is described in Algorithm 3.

c. Conceptual differences

In this subsection we discuss, from a conceptual point of view, the differences between AnDA and OI based on the formulations of these two algorithms. These differences are then assessed in sections 3 and 4 on numerical experiments.

The OI is a purely spatial-temporal interpolation method. The performance of OI completely relies on the choice of the static matrices $\mathbf{B}$ and $\mathbf{R}$. Hence, the interpolation does not account for the dynamics of the underlying state variable. As a consequence, the estimated posterior variance of the OI reanalysis shall only depend on the positions of observations and the physical locations of the state variables. On the other hand, AnDA automatically learns the dynamics from the catalog at every time step. Hence, the posterior variance of AnDA should be flow-dependent.

In operational usages of OI, it is usually not realistic to construct the full spatial-temporal climatological covariance. Hence, $\mathbf{B}$ is often assumed to be the tensor product of a spatial covariance matrix and a temporal correlation matrix. The temporal correlation matrix is uniquely determined by a scalar parameter L_t which defines the temporal correlation scale. Numerically, this artificial temporal correlation smooths out the temporal fluctuations of the reanalysis that have periods shorter than L_t. Hence, the OI should not be able to reconstruct the signal for modes of periods less

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

Algorithm 3 Analog forecast

Assume all the vectors are row vectors.

Input:$\mathbf{x}(t)$, k, $\mathscr{A}$, $\mathscr{S}$,

Output:$\mathbf{x}^f(t+1)$, $\mathbf{x}^p(t+1)$

Find the k analogs from $\mathscr{A}$ that are closest to $\mathbf{x}(t)$, denoted as $\mathbf{A}_1,...,\mathbf{A}_k$. And the distances d_i between $\mathbf{A}_i$ and $\mathbf{x}(t)$.

Find from $\mathscr{S}$ the successors of each $\mathbf{A}_i$, denoted as $\mathbf{S}_i$.

Define the weight w_i for each $\mathbf{A}_i$ based on the distance d_i.

$\bar{\mathbf{A}} \leftarrow \sum_{i=1}^{k} w_i \mathbf{A}_i$

$\mathbf{A}'_i \leftarrow \mathbf{A}_i - \bar{\mathbf{A}}$, let $\mathbf{A} = ((\mathbf{A}'_1)^\top,...,(\mathbf{A}'_k)^\top)^\top$

$\bar{\mathbf{S}} \leftarrow \sum_{i=1}^{k} w_i \mathbf{S}_i$

$\mathbf{S}'_i \leftarrow \mathbf{S}_i - \bar{\mathbf{S}}$, let $\mathbf{S} = ((\mathbf{S}'_1)^\top,...,(\mathbf{S}'_k)^\top)^\top$

Find the singular value decomposition of the matrix $\mathbf{A}$

Remove the small diagonal components of $\mathbf{S}$ and the corresponding columns to get $\mathbf{S}^{\text{red}}$. Remove the corresponding rows of $\mathbf{V}$ to get $\mathbf{V}^{\text{red}}$. Therefore, $\mathbf{A}^{\text{red}} = \mathbf{U}\mathbf{S}^{\text{red}}\mathbf{V}^{\text{red}}$ is an approximation of $\mathbf{A}$.

$\mathbf{W} \leftarrow \text{diag}(w_1,...,w_k)$

$\mathbf{C}_{xx} \leftarrow (\mathbf{A}^{\text{red}})^\top \mathbf{W} \mathbf{A}^{\text{red}}$

$\mathbf{C}_{xx2} \leftarrow (\mathbf{A}^{\text{red}})^\top \mathbf{W}^2 \mathbf{A}^{\text{red}}$

$\mathbf{C}_{xy} \leftarrow (\mathbf{A}^{\text{red}})^\top \mathbf{W} \mathbf{S}$

$\mathbf{M} \leftarrow (\mathbf{C}_{xx})^{-1} \mathbf{C}_{xy}$

$\eta_i \leftarrow \mathbf{S}_i - \bar{\mathbf{S}} - \mathbf{A}'_i \mathbf{M}$

$\mathbf{Q} \leftarrow \sum_{i=1}^{k} w_i (\eta_i^\top \eta_i)/(1 - tr(\mathbf{C}_{xx2}\mathbf{C}_{xx}^{-1}))$

$\mathbf{x}^p(t+1) \leftarrow (\mathbf{x}(t) - \bar{\mathbf{A}})\mathbf{M} + \bar{\mathbf{S}}$

$\mathbf{x}^f(t+1) \leftarrow \mathbf{x}^p(t+1) + \mathscr{N}(0, \mathbf{Q})$

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

than L_t. In contrast, AnDA does not have this limitation since the state variables are propagated under the dynamics learned from the catalog.

3. Application to the Lorenz-63 system

In this section, we compare the reanalysis means and variances produced by AnDA with those produced by OI, using the classic three-dimensional Lorenz-63 (L63) chaotic system (Lorenz 1963):

$$\begin{aligned} \frac{\mathrm{d}x_t}{\mathrm{d}t} &= 10\,(y_t - x_t)\,, \\ \frac{\mathrm{d}y_t}{\mathrm{d}t} &= x_t\,(28 - z_t) - y_t, \\ \frac{\mathrm{d}z_t}{\mathrm{d}t} &= x_t y_t - \frac{8}{3} z_t. \end{aligned} \tag{6}$$

The system is integrated with $dt = 0.01$ using the 4-th order Runge-Kutta method. The first component $x(t)$ is observed for every 10 time steps (i.e. $dt_{obs} = 0.1$), with an additive white Gaussian noise of variance $\mathbf{R} = 2$. After model spin-up, we first run the model for 10^3 time steps to generate the truth, and then we continue to run the model for 10^4 time steps to generate the catalog for AnDA. Our goal is to calculate the reanalysis of x together with its uncertainty estimate, based on the simulated observations. In this experiment we pretend that we have no knowledge of y and z. Therefore, we can not directly apply the L63 equations for forecasting, which is the scenario that AnDA is designed for.

a. Implementation of AnDA

Applying AnDA directly on the first L63 component can not lead to a good estimation. Indeed, if $x_t = a$, the intersection of the section $x = a$ and the L63 attractor has two branches, which is the case for a large proportion of possible values of a. Then whether x_{t+1} would be greater than or smaller than x_t depends on which branch the full state variable (x_t, y_t, z_t) lies on. Hence, it is

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

roughly equally likely for x_t to increase or decrease in the next time step. Therefore, we would not be able to have an informative prediction of x_{t+1} by merely looking at the analogs of x_t. A solution to this problem is to consider the time-delayed states $\mathbf{x}_t = (x_t, x_{t-\tau}, x_{t-2\tau})^\top$ for the implementation of AnDA. Experimentally, we find the optimal $\tau = 11$ value. Figure 1 shows the original attractor and the attractor of the time-delayed state variable. By using the time-delayed states as analogs, the details of the implementation of AnDA shall change correspondingly, which is explained in detail in the Appendix. We use an ensemble of size $N_e = 50$. At each time step, we apply analog forecasting separately to each ensemble member with $k = 50$, which is the parameter mentioned in step 1 of analog forecasting. For the Kalman smoother, we use $\mathbf{R} = 2$, which is the same as the observation error variance used to create the observations.

b. Implementation of OI

Since we only consider the first component x of the full system, we choose the following prior background covariance:

$$\mathbf{B}(x_{t_1}, x_{t_2}) = B_{11} \exp\{-|t_1 - t_2|^2 / L_t^2\}, \tag{7}$$

where B_{11} is the climatology covariance of x, which can be calculated from a long-time simulation. The parameters of OI, namely r and L_t as indicated by Eqs. (4,7), are tuned to guarantee that OI algorithm produces the minimal RMSE: here we set $r = 2$ and $L_t = 0.2$.

c. Comparison of mean estimates

Let $\hat{\mathbf{x}}$ be the reanalysis estimates of AnDA or OI, and $\mathbf{x}^{\text{true}}$ be the truth such that $\mathbf{x}^{\text{true}}$ and $\hat{\mathbf{x}}$ exist for $t_1, t_2, ..., t_T$. Suppose that $\hat{\mathbf{x}} = (\hat{x}_j)_{1,...,n_x}$ and $\mathbf{x}^{\text{true}} = (x_j^{\text{true}})_{1,...,n_x}$ are of dimension n_x (which equals 1 in the present L63 case), the RMSE of $\hat{\mathbf{x}}$ is then defined as:

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

$$\mathrm{RMSE} = \sqrt{\frac{1}{T}\frac{1}{n_x}\sum_{i=1}^{T}\sum_{j=1}^{n_x}\|\hat{x}_j(t_i) - x_j^{\mathrm{true}}(t_i)\|^2}. \tag{8}$$

Although the $\mathbf{x}_t$ we use for analog forecast with time-delayed states has three components, we only take the first component x_t to compute the RMSE. The time-delayed estimates (i.e. the second and the third components of the state reanalysis) are not used to evaluate the performance.

The RMSE for AnDA is 0.77, and the minimal (after tuning the parameters) RMSE for OI is 1.177. The top panel of Figure 2 shows the trajectory of the truth, the observation, and the reanalysis estimates of the L63 first component. The state reanalysis produced by OI apparently has large errors when the state is near the origin. In contrast, the trajectory of AnDA manages to reproduce the L63 dynamics even when the observation errors are large. In this experiment, we do not meet the curse of dimensionality, since we have 10^4 samples in the catalog while the Hausdorff dimension (Schleicher 2007) of the L63 attractor is around 2.06. Therefore, the dynamics represented by the analog forecast method approximates the true dynamics very well.

d. Comparison of estimated standard deviations

Another interesting way of comparing AnDA and OI is assessing the quality of the estimated standard deviation of the state reanalysis versus the true absolute error. Indeed, the absolute error directly quantifies how far the estimate is from the truth. However, the truth is usually unknown hence the absolute error is often not accessible. When this is the case, estimated standard deviations are often used as a reference to inform on the actual error of the state estimate. Hence, providing an estimated standard deviation that corresponds to the absolute error is a key feature

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

for a reconstruction method. These quantities are defined as follows

$$\text{stdev} = \sqrt{diag(\mathbf{P}^s)} \in \mathbb{R}^{n_x} \tag{9}$$

$$\text{abs error} = |\hat{\mathbf{x}} - \mathbf{x}^{\text{true}}| \in \mathbb{R}^{n_x}. \tag{10}$$

It is not surprising to see that the OI algorithm produces a periodic estimate of standard deviation (Fig. 2, bottom panel). Indeed, the estimated error is only based on the observation sampling. This is a strong limitation of OI. In contrast, the estimated standard deviation of AnDA is much more flow-dependent (Fig. 2, middle panel). The absolute error of AnDA increases each time the state variable is close to the bifurcation point or the furthest points of the two wings. At those times, the AnDA estimated standard deviation manages to inform on the error made as the complexity of the L63 dynamics renders the state estimation harder.

4. Application to the interpolation of along-track SSH

a. Targeted region and dataset:

In this section we test the OI and AnDA algorithms in an observing system simulation experiment (OSSE) aiming at interpolating along-track SSH onto gridded SSH maps. We focus here on a $10^\circ \times 10^\circ$ region in the eastern Gulf of Mexico (centering at $85^\circ W, 25^\circ N$, see Fig. 3). In terms of grid points, the region of interest is 41×41 large, including $n_x = 1353$ ocean grid points in total (the rest being land masses) thus giving the dimension of the state variable $\mathbf{x}$.

The ocean circulation in this region features the Loop Current (LC), an anti-cyclonic flowing meander entering the Gulf through the Yucatan channel (Yucatan current), and exiting along the southern tip of the Florida peninsula (Florida current). The Loop Current is known as an unstable system and episodically sheds large anti-cyclonic eddy rings of scale 200-400 km with periods ranging from about 100 to 450 days (see Fig. 3). The shedding of these Loop Current Eddies is

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

a complicated process as eddies can detach and reattach to the Loop Current, before propagating westward across the Gulf. SSH variability in the region is also related to smaller-size cyclonic eddies (80-120 km) that are observed moving along the outer edge of the LC (Loop Current frontal eddies, LCFEs), both on subannual and submonthly timescales, and to coastally-trapped waves that responds to wind variability, and especially to winter cold surges (see Jouanno et al. 2016, for a review).

We perform the OSSE using daily SSH maps from one of the OCCIPUT ensemble simulations (Penduff et al. 2014; Bessières et al. 2017). This is a regional North-Atlantic ocean/sea-ice 50-member ensemble simulation performed at eddy-permitting horizontal resolution (1/4°). After a common 20-year spinup, the 50 members are restarted from slightly perturbed initial conditions and forced over 20 years (1993-2012) with identical surface forcing. In the following, the SSH of the last year of the first ensemble member is taken as the ground truth. We then use the location of the real along-track AVISO observations available for 2004 (that include 4 satellites: TOPEX/Poseidon, GFO, Jason-1, ENVISAT), to generate our pseudo-observations by locally and linearly interpolating the truth along the observed tracks. No observation error is artificially added to the simulated observations (i.e. $\mathbf{R}_{\text{true}} = 0$).

The historical catalog from which AnDA learns the forecast model is thus made of the daily maps of SSH from the 19 remaining years of the 49 remaining ensemble members (meaning $19 \times 49 \times 365 = 339815$ daily SSH maps in total). As an element of comparison, the historical catalog in Lguensat et al. (2019) for a similar problem is 34 years of 3-day data (4017 SSH maps).

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

b. Implementation of AnDA

First we reduce the dimension of the state variable. We take the coefficients of the first 100 leading EOFs as the reduced state $\mathbf{x}^{red} \in \mathbb{R}^{100}$. In practice, we calculate the spatial climatology covariance $\mathbf{B}^{clim} \in \mathbb{R}^{1353\times1353}$ based on the OCCIPUT simulation:

$$\mathbf{B}^{clim} = \frac{1}{365000}\sum_{i_Y=1}^{20}\sum_{i_N=1}^{50}\sum_{t=1}^{365}(\mathbf{x}_{i_N,i_Y}(t)-\bar{\mathbf{x}})(\mathbf{x}_{i_N,i_Y}(t)-\bar{\mathbf{x}})^\top,$$

where $\mathbf{x}_{i_N,i_Y}(t)$ refers to the SSH on the t-th day of year i_Y of the i_N-th ensemble member. The EOFs (denoted by $\mathbf{e}_i$) are the eigenvectors of $\mathbf{B}^{clim}$:

$$\mathbf{B}^{clim}\mathbf{e}_i = \lambda_i\mathbf{e}_i,$$

for $i = 1,2,...,1353$. Then for a given state variable $\mathbf{x} \in \mathbb{R}^{1353}$, the reduced state is defined by

$$\mathbf{x}^{red} = (<\mathbf{x},\mathbf{e}_1>,<\mathbf{x},\mathbf{e}_2>,...,<\mathbf{x},\mathbf{e}_{100}>)^\top \in \mathbb{R}^{100}.$$

The first 100 EOFs explains more than 99% of the variance of SSH. This explained variance is stable over the whole time series of 20 years.

AnDA is implemented with respect to $\mathbf{x}^{red}$. Our catalog consists of the $\mathbf{x}^{red}$ that were calculated using 49 members (member 2 to member 50), from Year 1 to Year 19. Therefore, the catalog and the truth come from different members and years. By dimension reduction, the corresponding observation operator $\mathbf{H}^{red}$ is different from the original $\mathbf{H}$:

$$\mathbf{H}^{red}\mathbf{x}^{red} = \sum_{i=1}^{100} x_i^{red}\mathbf{H}\mathbf{e}_i.$$

And the corresponding observation error variance $\mathbf{R}_{obs}^{red}$ is no longer zero since the small components (i.e. $<\mathbf{x},\mathbf{e}_{101}>,...,<\mathbf{x},\mathbf{e}_{1353}>$) are missing in the reduced state variable. In this reduced space, covariance localization is implemented as $\mathscr{T}^{-1}(\mathscr{T}(\mathbf{P})\circ\mathbf{C}_{loc})$ in line 4 of Algorithm 2, where $\mathscr{T}$ transforms the covariances of $\mathbf{x}^{red}$ to the covariances of the original physical state $\mathbf{x}$.

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

We choose $N_e = 1000$ (ensemble size for data assimilation), $k = 1000$ (the parameter mentioned in Algorithm 3) and $\mathbf{R} = 4\text{cm}^2$. A different choice of analogs was made in Lguensat et al. (2019) where the analogs and successors were chosen to represent only the small-scale modes of the complete simulated SSH. The large-scale modes of SSH were first reconstructed using the OI method. Then the small-scale modes were reconstructed using AnDA. Although this space reduction strategy was shown to be promising, its success requires a catalog of high resolution SSH data which is not often available.

c. Implementation of OI

We considered the following background covariance matrix:

$$\mathbf{B}(x_{i,t_1}, x_{j,t_2}) \quad = \quad B_{ij} \exp\{-d_t^2 / L_t^2\}, \tag{11}$$

where $i, j = 1, 2, ..., \dim(\mathbf{x})$, d_{ij} refers to the physical distance between x_i and x_j, $d_t = |t_1 - t_2|$, $B_{ij} = Cov(x_i, x_j)$ refers to the spatial climatology covariance, and L_t is the scalar parameter defining the temporal scales of the covariance matrices.

The parameters B_{ij} are directly calculated from the SSH dataset and the parameter L_t is tuned so that the OI algorithm produces the minimal RMSE. Often in real applications, when the true spatial climatology covariances are not accessible, they are also parametrized and the background covariance matrix is approximated by $\mathbf{B}(x_i(t_1), x_j(t_2)) = \sqrt{B_{ii} B_{jj}} \exp\{-d_{ij}^2 / L_x^2 - d_t^2 / L_t^2\}$, with L_x a scalar parameter defining the spatial scales of the covariance matrices. However, for the sake of a fair comparison between OI and AnDA and since the simulated dataset in our experiments is large enough, we are able to estimate the spatial climatology covariances B_{ij} and fully compute Eq. (11). This formulation indeed yields the best results in the experiments of the present section (comparison not shown).

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

Note that, for the OI in the DUACS, the choice of the parameters L_x and L_t is usually made as a best global trade-off to achieve global resolution of the mesoscale features (e.g. Dibarboure et al. 2011; Pujol et al. 2012), and could, in principle, be better optimized in a specific regional context (hence the tuned OI in this study).

In this study, we also consider an OI optimized with conventional objective analysis (Le Traon et al. 1998) and here named OI_{COA}. The OI_{COA} experiment is used as a point of comparison in order to show that (i) an OI is difficult to tune (conventional objective analysis fails to do so) and (ii) an incorrectly tuned OI can lead to significant errors. The covariance function $\mathbf{B}^{COA}$ is chosen to be:

$$\mathbf{B}^{COA}(x_{i,t_1}, x_{j,t_2}) = \sqrt{B_{ii}B_{jj}}C_{ij}^{COA}\exp\{-d_t^2/L_t^2\}, \tag{12}$$

where

$$C_{ij}^{COA} = (1+\frac{\alpha d_{ij}}{L_x}+\frac{(\alpha d_{ij})^2}{6L_x^2}-\frac{(\alpha d_{ij})^3}{6L_x^3})\exp\{\frac{-\alpha d_{ij}}{L_x}\}, \tag{13}$$

with $\alpha = 3.34$. In Le Traon et al. (1998), the parameters are chosen to be $L_x = 150$ km, $L_t = 20$ days. We tune R so that the method produces the minimal RMSE based on the given L_x and L_t. A sensitivity test (not shown here) demonstrated that the difference between OI computed from Eq. (11) and OI_{COA} is mainly due to the difference in the parameter L_t. The correlation functions are also different but do not make a significant difference in our numerical results.

d. RMSE results

The RMSE values for SSH, vorticity and velocity reconstructed with the 3 methods (AnDA, OI and OI_{COA}) are summarized in Table 1. Here, the velocity refers to the two-dimensional vector of the geostrophic velocities which is defined as $(u,v) = (-\frac{\partial ssh}{\partial y}, \frac{\partial ssh}{\partial x})g/f$ and the vorticity is defined as $q = (\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y})g/f$, where $g = 9.81m/s^2$ is the gravity acceleration and f is the Coriolis force.

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

Table 1 shows that AnDA does as good as the best-tuned OI (i.e. tuned and optimized specifically for the region of interest) for these 3 variables, resulting in very similar RMSE values for the two methods in the full region of interest. In the case of SSH, the RMSE value for AnDA is smaller than the one for OI (1.40 cm vs 1.68 cm, resp.). However, this difference fades off when the RMSE is computed over the central region only, i.e. excluding coastal areas. In the following, we will show that this is due to the fact that AnDA can reconstruct the high-frequency SSH fluctuations of the coastal areas much better than OI. These SSH high-frequency fluctuations are likely related to the coastally-trapped waves responding to winter wind storm surges as mentioned in Jouanno et al. (2016).

It is also clear from Table 1 that OI_{COA} is systematically less accurate than AnDA and OI in terms of RMSE. The time series of the reconstructed SSH at two example grid points, displayed in Fig. 4, provide an illustration to why this is the case. With parameter L_t set to 20 days for the temporal correlation scale of OI_{COA}, the reconstructed SSH misses the high-fluctuations of the signal. These high-frequency fluctuations are particularly strong near the Florida coast (bottom panel), while in the Loop Current (top panel), the large amplitude fluctuations appear to be of monthly and sub-annual timescales as they are associated with the fluctuations of the LC meander and LCE shedding. On the other hand, the tuning of the best-tuned OI with $L_t = 6$ days results in a better behavior of the reconstructed SSH in the high-frequencies. We quantify this further in the following with a dedicated temporal spectral analysis. At this point, we wish to emphasize the fact that AnDA is as accurate as the best-tuned OI, without the need to explicitly tune the parameters L_x and L_t. In AnDA, the information is implicitly provided by the historical catalog. It should be reminded, however, that these results are produced in the context of an OSSE with pseudo-observations derived from the simulated truth, and so the historical catalog from which

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

AnDA learns is fully consistent with those observations. We reserve for future investigations the case where the catalog and the truth come from different sources.

An additional sensitivity test has been performed in order to assess the impact of the catalog size on the reconstruction performances. Three other catalog sizes have been implemented: using 1 member (19×365 daily SSH maps), 20 members (20×19×365 daily SSH maps) and 30 members (30×19×365 daily SSH maps) of the 19 year OCCIPUT ensemble and compared to the current catalog using 49 members (49×19×365 daily SSH maps). The resulting SSH reconstruction RMSE are respectively: 1.82, 1.46, 1.45 and 1.40 cm. As expected, the performance of AnDA are improved by a larger catalog. However, the dependence is not linear and the difference between using 20 members, 30 members and 49 members is relatively small. In fact, both the 20 member catalog and the 30 member catalog also lead to smaller RMSE than OI.

e. Temporal spectral results

The top panel of Figure 5 shows the temporal power spectral densities (PSD) averaged over the entire domain for the reconstructed SSH with the three methods and for the true SSH. The PSD of the three reconstructed signals are very close to the PSD of the truth at timescales longer than about 30 days, confirming that all three methods produce equivalent energy reconstructions of the monthly-to-sub-annual fluctuations. But at higher frequencies, we find that only the PSD for the AnDA-reconstructed SSH stays close to the truth. A drop-off in the PSD is clearly seen for the OI- and OI_{COA}-reconstructed SSH at approximately 6 and 20 days respectively which is consistent with the values set for L_t in each case.

We also check the noise-to-signal ratio between the reconstructed signals and the truth (Fig. 5-bottom panel). For this purpose, and following Dufau et al. (2016) and Ballarotta et al. (2019), we

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

compute the spectral noise-to-signal ratio as:

$$R = \frac{1 - \mathrm{PSD}_{\mathrm{Error}}}{\mathrm{PSD}_{\mathrm{Truth}}},$$

where $\mathrm{PSD}_{\mathrm{Error}}$ is the PSD of the difference between the reconstructed SSH and the truth, and $\mathrm{PSD}_{\mathrm{Truth}}$ is the PSD of the true signal. This metric provides a measure of the coherence between the two signals that takes into account differences in both amplitude and phase (Ballarotta et al. 2019). Figure 5-bottom panel shows that the spectral noise-to-signal ratio for AnDA remains far above 0.5 down to time scales of $\sim$5 days, which confirms that a good coherence exists between the AnDA-reconstructed and the true SSH. The coherence is not as good for the two reconstructed SSH signals of OI and OI_{COA}. For the best-tuned OI, R drops below 0.5 at time scales of $\sim$15 days, even if the PSD drops off only around 6 days. In other words, the OI method manages to reconstruct enough energy at high-frequencies (although not below 6 days) yet fails to produce a coherent signal at those scales. As for OI_{COA}, both PSD and R drop off at time-scales of about 25 days.

We thus find confirmation in Fig. 5 that AnDA is able to reconstruct an SSH signal with good coherence to the truth at higher frequencies than OI, and so even when AnDA is compared with the OI specifically tuned for the region of interest. This is consistent with what we had already pointed out from the example grid-point timeseries in Fig. 4. As already discussed, in the domain we examine here, sub-monthly fluctuations are the strongest in coastal regions because of the response to wind bursts (Jouanno et al. 2016). Operational systems such as DUACS, based on OI, are not able to capture well those fast fluctuations, but can partially get around this limitation by using additional products such as the AVISO-DAC (Dynamic Atmosphere Correction) to propose an a posteriori correction for the missed and aliased part of the signal corresponding to the dynamical high-frequency ocean response to wind and pressure forcing (e.g. Ponte and Ray 2002). Note

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

that this correction, however, is only based on this specific source of high-frequency fluctuations, while our study shows that a method such as AnDA is able to capture high-frequency signals originating from all kind of sources (to the extent that the fluctuations are well represented in the historical catalog). We find indeed that the spectral results shown in Fig. 5 remain robust also when restricting the area of the spectral analysis to the central ocean-only region (not shown here), meaning that AnDA is able to capture high-frequency fluctuations of any kind, and not only the coastally-trapped waves.

f. Estimated standard deviation results

Another interesting result of this study is that, consistently with the L63 application in section 3, AnDA produces a more informative estimated standard deviation. Indeed, it has similar spatial patterns as the absolute error (i.e. the difference with the true signal), and does not only depend on the tracks of the observations to interpolate. This is illustrated in Fig. 6 where estimated standard deviation and absolute error are averaged over a one month period on two example dates (March 8th, 2004 and September 8th, 2004) for AnDA, OI, and OI_{COA}. For visual purposes, we show the monthly averaged distribution of the absolute error as it presents a clearer flow-dependent feature than the daily distribution.

Figure 6 shows that the absolute error of AnDA on March 9 and September 9 is smaller than that of OI and OI_{COA}, especially along the Florida coast and in the loop current. It means that on that dates, the AnDA-reconstructed SSH is closer to the truth, which is consistent with time series given in Fig. 4. For instance, on September 9, the absolute errors concentrate near the anti-cyclonic flowing meander. And it is clear that in this region, the absolute error of AnDA is smaller than that of OI and OI_{COA}.

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

Figure 6 also illustrates the fact that the estimated standard deviation for OI depends on the satellite observation sampling (here, along-tracks) and on the background error covariance matrix **B**. This is consistent with the results given in Fig. 2 (bottom panel) in the case of the L63 system. The estimated standard deviation for OI is thus non informative. In contrast, the estimated standard deviation of AnDA does not only depend on the observation sampling but also on the flow. Therefore, its pattern is more correlated to the absolute error (see top and bottom left panels of each snapshot in Fig. 6).

5. Conclusion

This paper reviews the algorithms of analog data assimilation (AnDA) and optimal interpolation (OI), and presents the numerical results of interpolation with the Lorenz-63 (L63) system and with simulated sea-surface height (SSH) data. Our comparison of AnDA and OI mainly focuses on the root-mean-square error (RMSE) of the state estimate, the estimated standard deviation and the temporal spectra of the reconstructed states. In order to achieve a fair comparison, we carefully tune the parameters of OI so that the RMSE is the most reduced. As a reference we also present the numerical results of OI for a classical but suboptimal set of parameters (labeled OI_{COA}) in the experiments with SSH data. This setting corresponds to the seminal work described in Le Traon et al. (1998).

In the tests with the L63 model, a case where we do not meet the curse of dimensionality, we show that AnDA produces more realistic interpolated trajectories, especially when the true state is near the center of the system attractor (see Fig. 2 top panel). Meanwhile, the standard deviation estimated by AnDA is highly correlated with the absolute error, which is unknown in practice, and is hence much more informative. On the other hand, the standard deviation estimated by OI is

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

uncorrelated with the absolute error (see Fig. 2 middle and bottom panels) and only depends on the background and observation terms.

In the tests with simulated SSH data, AnDA and OI produce comparable RMSE for the daily SSH estimates (Table 1). However, only the interpolation using AnDA captures well the high-frequency fluctuations, including those generated in the coastal regions in response to winter wind bursts (Fig. 4). We show that the reconstructed temporal spectra of AnDA is also more consistent to that of the truth, in terms of energy and coherence, both at large and small time scales. In contrast, the OI-reconstructed temporal spectra suffers a significant loss of energy and is incoherent with the truth at small time scales (see Fig. 5). Moreover, the standard deviation estimated by AnDA is once again more informative. Indeed, compared to OI results, the AnDA estimated standard deviation is flow-dependent, evolving in space and time, and has a significant correlation with the absolute error (see Fig. 6).

To summarize, AnDA and OI are interpolation methods with slightly different formulations. In the case of OI, parameters controlling spatial-temporal variability and levels of noise are prescribed by the user. The optimization process of these parameters is time demanding, especially for large systems. Instead, AnDA is using analogs and these parameters are adaptively learned from a catalog of data, which needs to be as rich as possible. In one sense, the construction of the catalog in AnDA is time demanding but once it is created, this procedure is very convenient as it does not need additional tuning. In terms of interpolation results, AnDA and OI differ from their mean and standard deviation estimates. Regarding the mean estimate, AnDA, based on a catalog of numerical simulations, creates realistic trajectories which capture fast and slow fluctuations at the same time. Instead, OI is linearly interpolating the observations with static parameters, which makes OI incapable of capturing time scales that are smaller than the temporal correlation parameter. Regarding the standard deviation, OI can only estimate a standard deviation that is dependent

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

on the background and observation error covariances. AnDA is producing much more realistic standard deviation estimates, correlated with the absolute error of interpolation. This means that AnDA is able to detect when and where the interpolation is relevant or not. This point is crucial for the quantification of the uncertainty in the interpolation.

Our study demonstrates the potentiality of using AnDA as an alternative method to OI for the interpolation of along-track satellite observations. As the first step of demonstration, we have investigated for this study pure "twin" experiments, where the pseudo-observations and the AnDA historical catalog came from the same source (i.e. were fully consistent), and where a comparison to the known true SSH is possible. These twin experiments lead to encouraging results for AnDA, and call for future work to further test AnDA in the context of realistic operational applications. Future work will need to address several questions. First, are the good performances of AnDA confirmed when the historical catalog and the along-track observations do not come from the same source ? In other words, a realistic experimental study should be performed with real observations or at least artificial observations extracted from an entirely distinct numerical simulation. Second, is the AnDA method applicable to other regions and/or to global scale ? The current implementation is sufficient (technically) and can be straightforwardly applied to any other region of similar size as the Gulf of Mexico with no additional implementation difficulties. In this case, the creation of a new catalog will require new model data which can be costly unless, like in the present study with OCCIPUT data, the catalog is based on data that are available globally. The good performance of AnDA at regional scale (as shown here for the Gulf of Mexico) should then be confirmed in other regions under the condition that a computationally reasonable number of EOF is enough to capture the dynamics of that region. In other words, for this specific implementation of AnDA to work well, the energy distribution of the signal's EOF decomposition must present a small tail. For the same reason, the EOF-based AnDA implementation is most likely not suited (as

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

is) for global scale applications. Being able to maintain a relatively small and detailed catalog is crucial to ensure a successful analog research. The EOF decomposition (with a computationally reasonable number of EOFs) would fail to capture the detailed SSH signal in a larger region and even more so at global scale. This restriction does pose an important challenge to a global scale implementation of AnDA. However, a solid lead to extend the AnDA implementation to global scale has recently been developed. This implementation is a mixture of EOF-based AnDA implementation, as used in the present paper, and patched-based AnDA implementation, as described in Lguensat et al. (2019). This new implementation is currently under scrutiny and is already showing promising results. Finally, what is the computational cost of AnDA in comparison to OI ? For the moment, the computational cost of AnDA is much larger than OI but, as already mentioned, a strong argument for AnDA is that the method does not require as much tuning as OI. Moreover, in a realistic setting, the tuning of OI is not only complicated and time-consuming but the tuning optimality can not be guaranteed. Although, these considerations are obviously hard to quantify, a study should be conducted in where the computational efficiency of both OI and AnDA codes have been optimized. Also, in order to appropriately quantify the tuning efforts, the study should be taking into account the entire mapping production chain. A logical next step for AnDA would hence be to implement a comparative study in a realistic altimetric mapping production context in close collaboration with operational institutions.

Acknowledgments. This work has been carried out as part of the Copernicus Marine Environment Monitoring Service (CMEMS) 3DA project. CMEMS is implemented by Mercator Ocean in the framework of a delegation agreement with the European Union. Sammy Metref was funded by ANR through contract number ANR-17-CE01-0009-01. The ensemble simulation dataset used in this study was produced as part of the OCCIPUT project (http://meom-

group.github.io/projects/occiput/) funded by the French Agence Nationale de la Recherche (ANR) through contract ANR-13-BS06-0007-01, and further supported by the PIRATE project funded by the Centre National d'Études Spatiales (CNES) through the Ocean Surface Topography Science Team (OST/ST). The original OCCIPUT dataset is available upon request (contact: thierry.penduff@cnrs.fr)

APPENDIX

Time-delayed analog forecast

A key aspect of analog forecasting is how to choose the analogs. On one hand, the analogs need to be informative, meaning that Motivated by the mathematical theory established by Takens (1981) stating that, under certain conditions, the attractor of the original system can be embedded into the space of lagged partial state variables, we also consider using time-delayed states as the extended state variable. For the numerical experiment with Lorenz-63 system, our state estimate at time t is the 3-dimensional vector $\mathbf{x}^{\text{lag}}(t) = (x_t, x_{t-\tau}, x_{t-2\tau})^\top$, where x_t is the first component of the Lorenz-63 full state at time t and τ is a prescribed time gap. The value of τ is discussed in section 3.a. For each t, although $x(t)$ is represented in $\mathbf{x}^{\text{lag}}(t)$,$\mathbf{x}^{\text{lag}}(t+\tau)$,and $\mathbf{x}^{\text{lag}}(t+2\tau)$, we do not update $\mathbf{x}^{\text{lag}}(t)$,$\mathbf{x}^{\text{lag}}(t+\tau)$,$\mathbf{x}^{\text{lag}}(t+2\tau)$ at the same time. In other words, at the forecasting step at time $t-1$ or at the data assimilation step at time t, only $\mathbf{x}^{\text{lag}}(t)$ would be updated.

However, we do not apply time-delayed states in the experiment with SSH data since experimentally we do not find improvement of the quality of reanalysis.

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

References

Ballarotta, M., and Coauthors, 2019: On the resolutions of ocean altimetry maps. *Ocean Science*, **15 (4)**, 1091–1109, doi:10.5194/os-15-1091-2019, URL https://www.ocean-sci.net/15/1091/2019/.

Beckers, J. M., and M. Rixen, 2003: EOF calculations and data filling from incomplete oceanographic datasets. *J. Atmos. Oceanic Technol.*, **20**, 1839–1856.

Bentley, J. L., 1975: Multidimensional binary search trees used for associative searching. *Communications of the ACM*, **18 (9)**, 509–517.

Bessières, L., and Coauthors, 2017: Development of a probabilistic ocean modelling system based on NEMO 3.5: application at eddying resolution. *Geoscientific Model Development*, **10**, 1091–1106.

Compo, G. P., and Coauthors, 2011: The Twentieth Century Reanalysis Project. *RMetS*, **137**, 1–28.

Dibarboure, G., M.-I. Pujol, F. Briol, P. Y. L. Traon, G. Larnicol, N. Picot, F. Mertz, and M. Ablain, 2011: Jason-2 in duacs: Updated system description, first tandem results and impact on processing and products. *Marine Geodesy*, **34 (3-4)**, 214–241, doi:10.1080/01490419.2011.584826.

Dufau, C., M. Orsztynowicz, G. Dibarboure, R. Morrow, and P.-Y. Le Traon, 2016: Mesoscale resolution capability of altimetry: Present and future. *Journal of Geophysical Research: Oceans*, **121 (7)**, 4910–4927, doi:10.1002/2015JC010904, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015JC010904.

Durand, M., L. Fu, D. P. Lettenmaier, D. E. Alsdorf, E. Rodriguez, and D. Esteban-Fernandez, 2010: The surface water and ocean topography mission: Observing terrestrial surface water and

oceanic submesoscale eddies. *Proceedings of the IEEE*, **98 (5)**, 766–779, URL https://doi.org/10.1109/JPROC.2010.2043031.

Fablet, R., J. Verron, B. Mourre, B. Chapron, and A. Pascual, 2018a: Improving mesoscale altimetric data from a multitracer convolutional processing of standard satellite-derived products. *IEEE Transactions on Geoscience and Remote Sensing*, **56 (5)**, 2518–2525.

Fablet, R., P. Viet, R. Lguensat, P.-H. Horrein, and B. Chapron, 2018b: Spatio-temporal interpolation of cloudy sst fields using conditional analog data assimilation. *Remote Sensing*, **10 (2)**, 310, doi:10.3390/rs10020310, URL http://dx.doi.org/10.3390/rs10020310.

Fu, L.-L., and R. Ferrari, 2008: Observing oceanic submesoscale processes from space. *Eos, Transactions American Geophysical Union*, **89 (48)**, 488–488, doi:10.1029/2008EO480003, URL https://doi.org/10.1029/2008EO480003.

Gandin, L. S., 1965: Objective analysis of meteorological fields. *Israel Program for Scientific Translations*.

Gaspari, G., and S. E. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Q. J. R. Meteorol. Soc.*, **125**, 723–757.

Houtekamer, P. L., and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Monthly Weather Review*, **126**, 796–811.

Jouanno, J., J. Ochoa, E. Pallàs-Sanz, J. Sheinbaum, F. Andrade-Canto, J. Candela, and J.-M. Molines, 2016: Loop current frontal eddies: Formation along the campeche bank and impact of coastally trapped waves. *Journal of Physical Oceanography*, **46 (11)**, 3339–3363, doi:10.1175/JPO-D-16-0052.1.

Le Traon, P.-Y., F. Nadal, and N. Ducet, 1998: An improved mapping method of multisatellite altimeter data. *J. Atmos. Ocean. Technol*, **15**, 522–534.

Le Traon, P. Y., and Coauthors, 2019: From observation to information and users: The copernicus marine service perspective. *Frontiers in Marine Science*, **6**, 234, doi:10.3389/fmars.2019.00234, URL https://doi.org/10.3389/fmars.2019.00234.

Lguensat, R., P. Tandeo, P. Ailliot, M. Pulido, and R. Fablet, 2017: The Analog Data Assimilation. *Monthly Weather Review*, **145 (10)**, 4093–4107.

Lguensat, R., P. H. Viet, M. Sun, G. Chen, F. Tian, B. Chapron, and R. Fablet, 2019: Data-driven Interpolation of Sea Level Anomalies using Analog Data Assimilation. *Remote Sensing*, **11 (7)**, 858.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *journal of the atmospheric sciences*, **20**, 130–141.

Minamide, M., and F. Zhang, 2017: Adaptive Observation Error Inflation for Assimilating All-Sky Satellite Radiance. *Monthly Weather Review*, **145**, 1063–1081.

Miyoshi, T., E. Kalnay, and H. Li, 2013: Estimating and including observation-error correlations in data assimilation. *Inverse Problems in Science and Engineering*, **21 (3)**, 387–398.

Penduff, T., and Coauthors, 2014: Ensembles of eddying ocean simulations for climate. *CLIVAR Exchanges*, **65 (19-2)**.

Ponte, R. M., and R. D. Ray, 2002: Atmospheric pressure corrections in geodesy and oceanography: A strategy for handling air tides. *Geophysical Research Letters*, **29 (24)**, 6–1–6–4, doi:10.1029/2002GL016340, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2002GL016340.

Pujol, M.-I., G. Dibarboure, P.-Y. Le Traon, and P. Klein, 2012: Using high-resolution altimetry to observe mesoscale signals. *Journal of Atmospheric and Oceanic Technology*, **29 (9)**, 1409–1416, doi:10.1175/JTECH-D-12-00032.1.

Pujol, M.-I., Y. Faugère, G. Taburet, S. Dupuy, C. Pelloquin, M. Ablain, and N. Picot, 2016: Duacs dt2014: the new multi-mission altimeter data set reprocessed over 20 years. *Ocean Science*, **12 (5)**, 1067–1090, doi:10.5194/os-12-1067-2016, URL https://www.ocean-sci.net/12/1067/2016/.

Schleicher, D., 2007: Hausdorff dimension, its properties, and its surprises. *The American Mathematical Monthly*, **114 (6)**, 509–528, doi:10.1080/00029890.2007.11920440, URL https://doi.org/10.1080/00029890.2007.11920440, https://doi.org/10.1080/00029890.2007.11920440.

Takens, F., 1981: Detecting strange attractors in turbulence. *Dynamical systems and turbulence, Warwick 1980*, D. A. Rand, and L.-S. Young, Eds., Springer-Verlag Berlin Heidelberg, Vol. 898, 366–381, doi:10.1007/BFb0091903.

Tandeo, P., P. Ailliot, M. Bocquet, A. Carrassi, T. Miyoshi, M. Pulido, and Y. Zhen, 2018: Joint estimation of model and observation error covariance matrices in data assimilation: a review. *arXiv preprint arXiv:1807.11221*.

Tandeo, P., P. Ailliot, J. J. Ruiz, A. Hannart, B. Chapron, R. Easton, and R. Fablet, 2015: Combining analog method and ensemble data assimilation: application to the Lorenz-63 chaotic system. *Machine Learning and Data Mining Approaches to Climate Science*, 3–12.

Ubelmann, C., P. Klein, and L. L. Fu, 2015: Dynamic interpolation of sea surface height and potential applications for future high-resolution altimetry mapping. *Journal of Atmospheric and Oceanic Technology*, **32 (1)**, 177–184, doi:10.1175/JTECH-D-14-00152.1.

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

Wu, Z., 1995: Compactly supported positive definite radial functions. *Advances in Computational Mathematics*, **(1995)**.

Zhen, Y., P. Tandeo, S. Leroux, S. Metref, T. Penduff, and J. L. Sommer, 2019: 3da code and data. Zenodo, URL https://doi.org/10.5281/zenodo.3559784, doi:10.5281/zenodo.3559784.

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

LIST OF TABLES

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

Variable	Domain	AnDA	OI	OI_{COA}
SSH (in cm)	Full domain	1.40	1.68	2.18
	No coast (*)	1.33	1.39	1.65
	FY Coasts	1.76	2.95	3.73
Velocity (in $cm.s^{-1}$)	Full domain	5.68	5.57	7.44
	No coast (*)	6.35	6.28	7.78
	FY Coasts	3.33	5.05	6.99
Vorticity ($(100s)^{-1}$)	Full domain	0.220	0.212	0.293
	No coast (*)	0.226	0.242	0.292
	FY Coasts	0.101	0.167	0.249

TABLE 1. Summary of the RMSE values obtained with the 3 methods for year 2004 (06-01-2004 to 31-12-2004): AnDA, OI and OI_{OI_COA}, for SSH (in cm), geostrophic velocity (in $cm.s^{-1}$ and vorticity $((100s)^{-1})$, computed over the full domain, in the central region only (*), i.e. excluding the coastal areas (longitude 83.75°W-90°W; latitude 23.78°N-27.13°N), and in the Florida and Yucatan coastal area

.

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

LIST OF FIGURES

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.

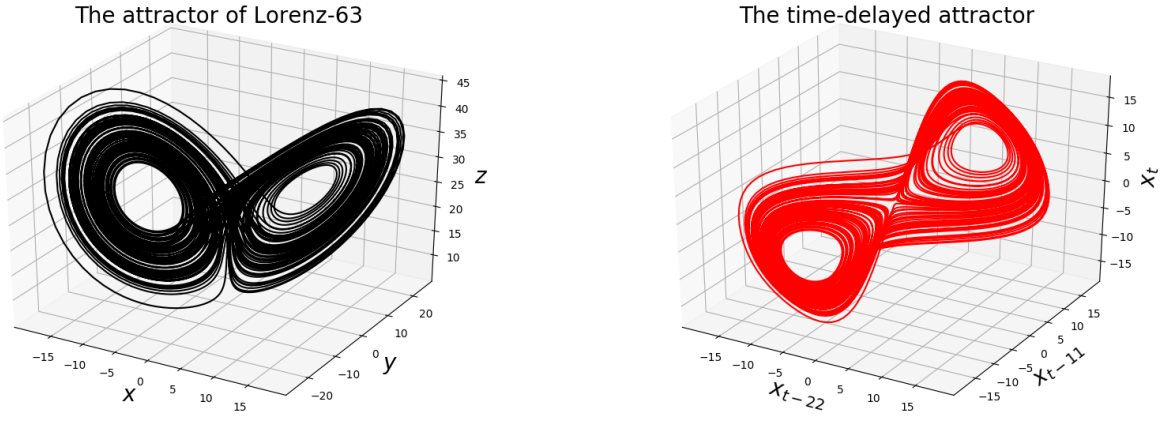
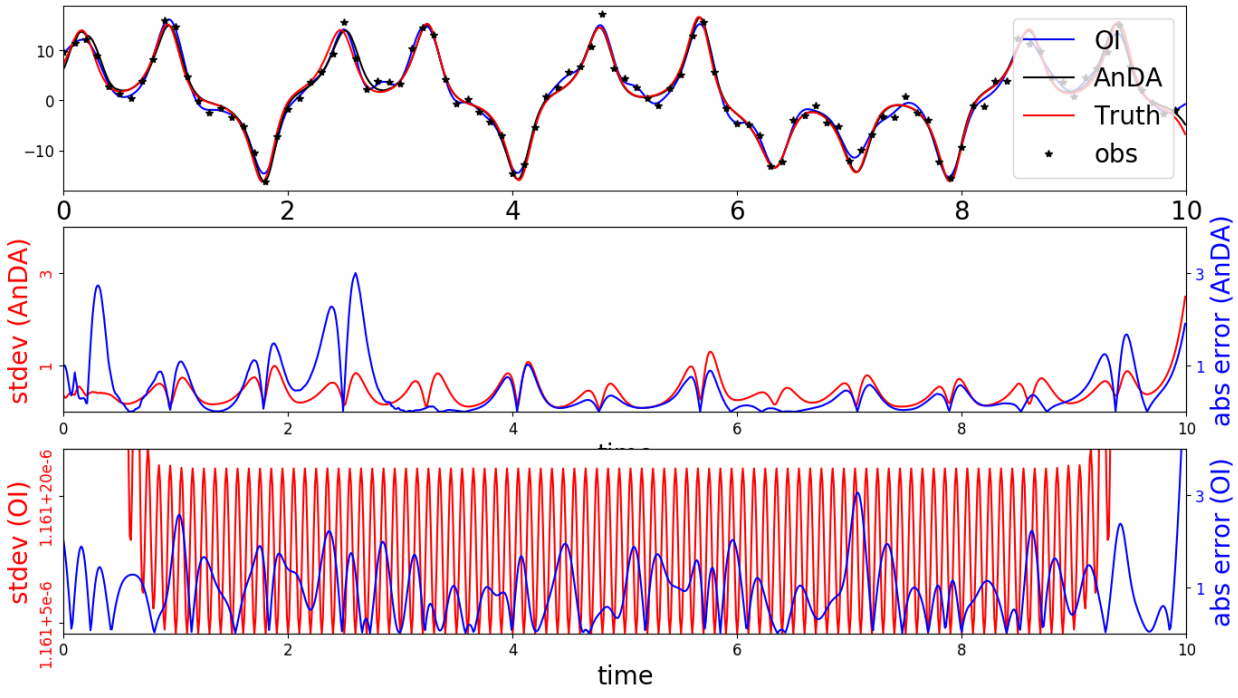


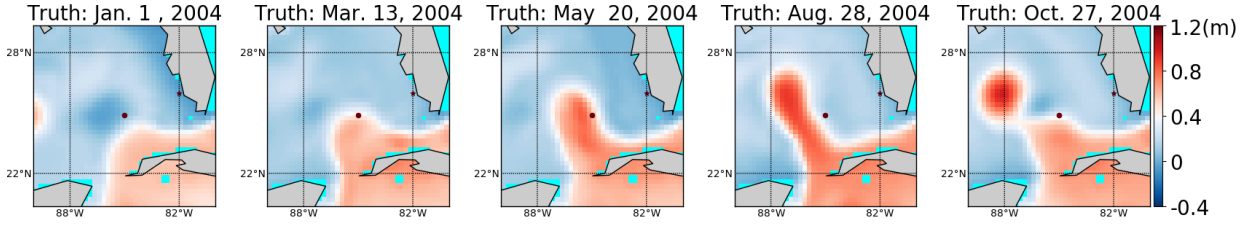
FIG. 1. The attractors of the orginal state variable and the time-delayed state variable of L63.

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.



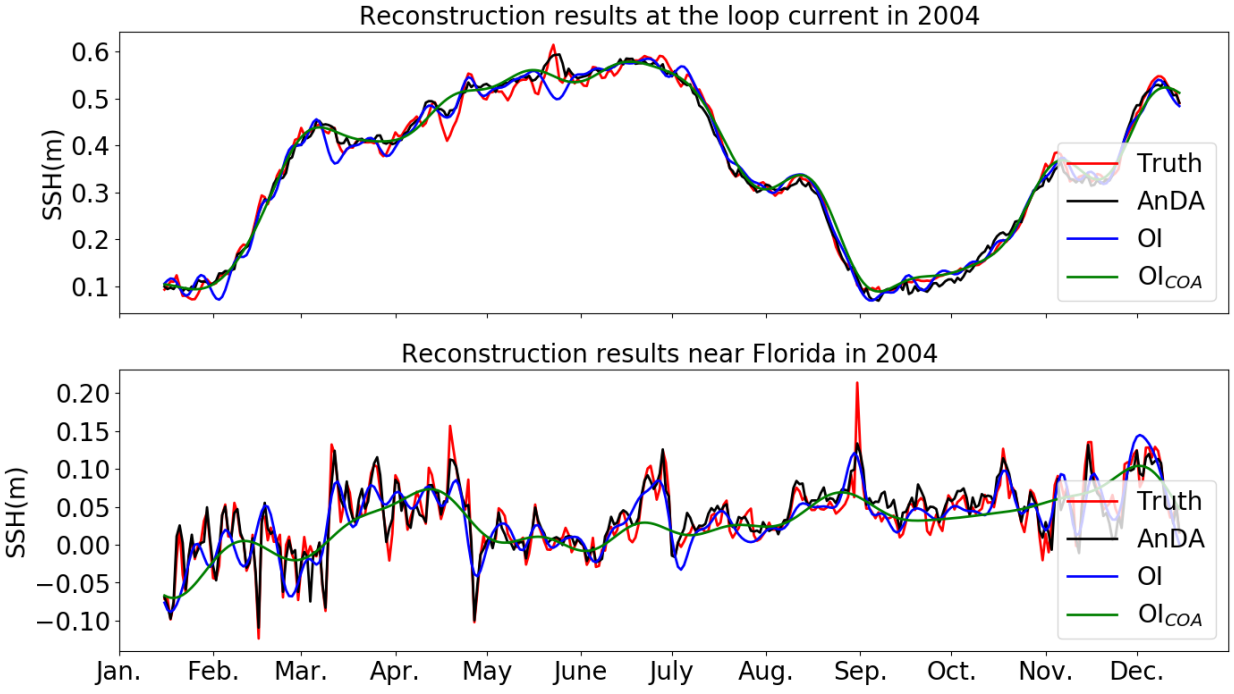
719 FIG. 2. (Top) The trajectory of the truth, the observations, the AnDA and OI estimates. The RMSE of AnDA
720 estimates and OI estimates are 0.77 and 1.177, respectively. (Middle) Estimated reanalysis standard deviation
721 and absolute error of reanalysis estimate for AnDA. The estimated standard deviation is strongly correlated to
722 the absolute error. (Bottom) estimated reanalysis standard deviation and absolute error of reanalysis estimate
723 for OI. In this example, the estimated standard deviation is periodic since it only depends on the observation
724 frequency and the magnitude of **R**.

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.



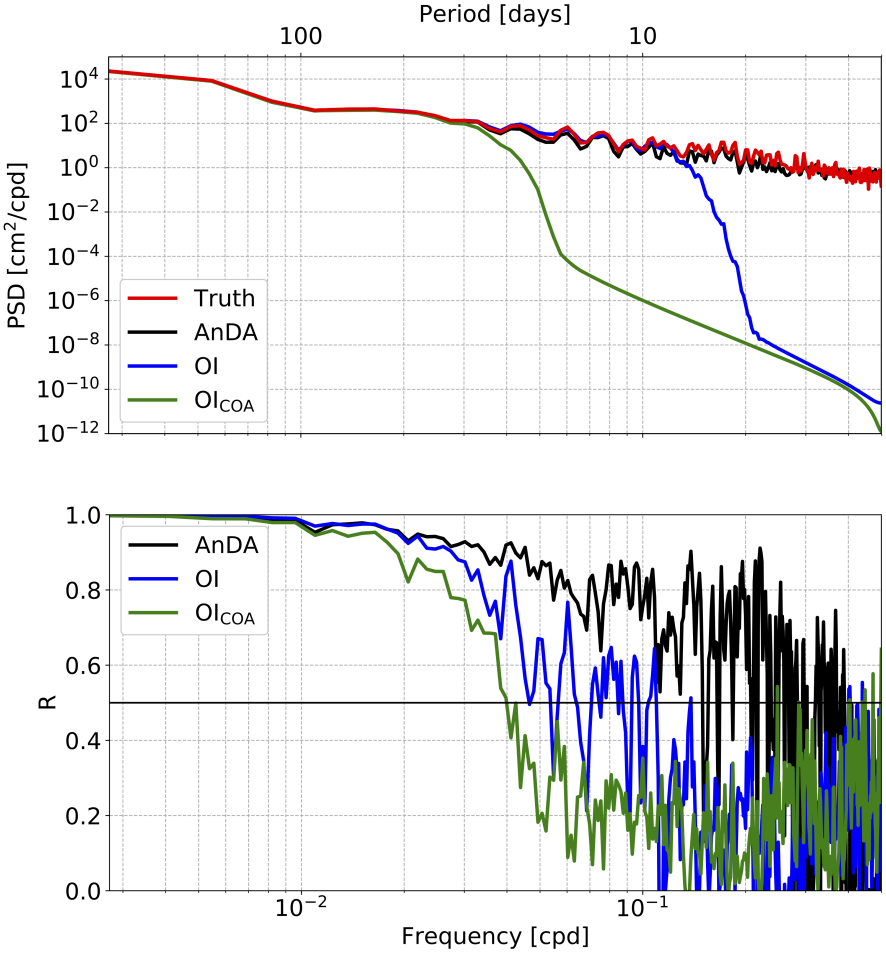
725 FIG. 3. Snapshots of the "true" SSH in the region of interest on different days of year 2004, featuring the
726 formation and shedding of a "Loop Current Eddy". The SSH here comes from the OCCIPUT ensemble simula-
727 tion (see text). The two symbols on the maps mark the location of the Loop-Current and the Florida-Coast grid
728 points, respectively at 85°W,25°N and at 82°W,26.03°N.

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.



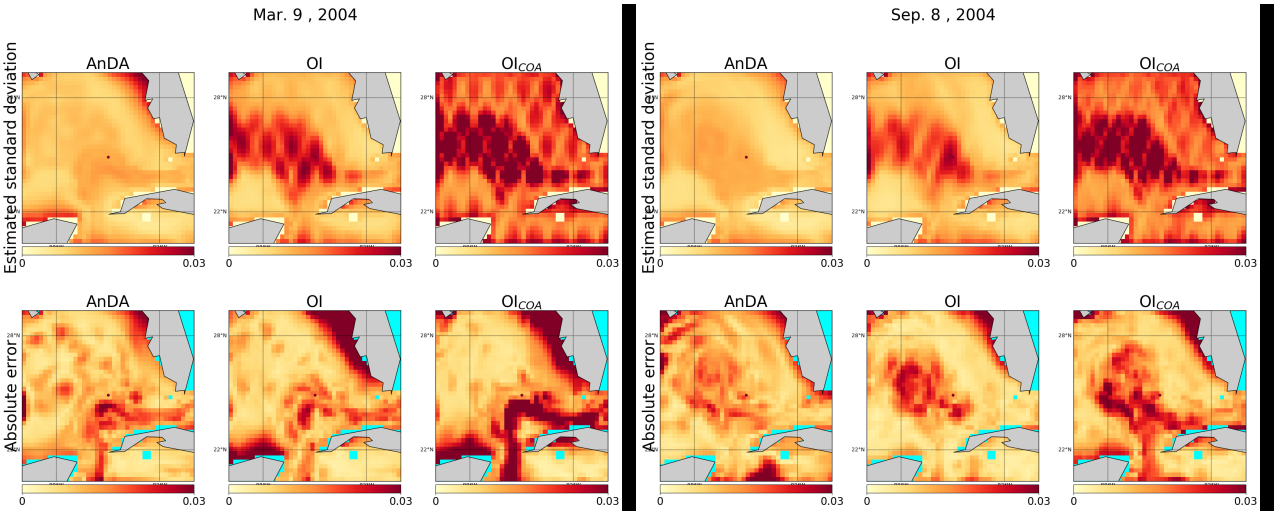
729 FIG. 4. Timeseries of the reconstructed daily SSH for year 2004 at the two gridpoints marked on the maps in
730 Fig. 3: in the Loop Current (85°W,25°N) and near the Florida coast (82°W,26.03°N). The reconstructed SSH is
731 shown for AnDA, OI and OI$_{COA}$ and compared with the true SSH.

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.



732 FIG. 5. (Top) Temporal power spectral density (PSD) of the reconstructed SSH (AnDA, OI, OI_{COA}) and true
733 SSH. (Bottom) Temporal signal-to-noise ratio (R) measuring the temporal coherence of each of the reconstructed
734 SSH (AnDA, OI, OI_{COA}) with the true SSH. Both PSD and R are averaged over the entire domain. Both panels
735 share the same x-axis in log scale for temporal frequency (cycles per day: cpd). The tick labels on the top axis
736 give the corresponding periods in days.

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.



737 FIG. 6. Monthly averages of estimated standard deviation and absolute error centered on March 8th, 2004 and
738 September 8th, 2004. The $\mathbf{P}^s$ produced by OI (upper middle panel for each month) and OI_{COA} (upper right panel
739 for each month) only depends on the tracks of satellite altimetry and the background covariance $\mathbf{B}$. Therefore,
740 the estimated standard deviation does not seem relevant to approximate the absolute error (lower middle and
741 lower right panels for each month). On the other hand, the estimated standard deviation produced by AnDA is
742 flow dependent (upper left panel for each month) and closer to the absolute error.

Accepted for publication in *Journal of Atmospheric and Oceanic Technology.* DOI 10.1175/JTECH-D-20-0001.1.