Data-driven Interpolation of Sea Level Anomalies using Analog Data Assimilation

Redouane Lguensat, Phi Huynh Viet, Miao Sun, Ge Chen, Tian Fenglin, Bertrand Chapron, Ronan Fablet

Data-driven Interpolation of Sea Level Anomalies using Analog Data Assimilation

Redouane Lguensat[a,*], Phi Huynh Viet[a], Miao Sun[b], Ge Chen[b], Tian Fenglin[b], Bertrand Chapron[c], Ronan Fablet[a]

[a]*IMT Atlantique, Lab-STICC, Université Bretagne Loire*
[b]*Department of Marine Technology, College of Information Science and Engineering, Ocean University of China, Qingdao, China*
[c]*Laboratoire d'Océanographie Physique et Spatiale, Centre de Brest, IFREMER, Plouzané, France*

Abstract

Despite the well-known limitations of Optimal Interpolation (OI), it remains the conventional method to interpolate Sea Level Anomalies (SLA) from altimeter-derived along-track data. In consideration of the recent developments of data-driven methods as a means to better exploit large-scale observation, simulation and reanalysis datasets for solving inverse problems, this study addresses the improvement of the reconstruction of higher-resolution SLA fields using analog strategies. The reconstruction is stated as an analog data assimilation issue, where the analog models rely on patch-based and EOF-based representations to circumvent the curse of dimensionality. We implement an Observation System Simulation Experiment in the South China sea. The reported results show the relevance of the proposed framework with a significant gain in terms of root mean square error for scales below 100km. We further discuss the usefulness of the proposed analog model as a means

*corresponding author, email: redouane.lguensat@imt-atlantique.fr

to exploit high-resolution model simulations for the processing and analysis of current and future satellite-derived altimetric data.

Keywords: Analog Data Assimilation, Sea Level Anomaly, Sea Surface Height, Interpolation, Data-driven methods

1. Introduction

The past twenty years have witnessed a deluge of ocean satellite data, such as sea surface height, sea surface temperature, ocean color, ocean current, sea ice, etc. This has helped building big databases of valuable information and represents a major opportunity for the interplay of ideas between ocean remote sensing community and the data science community. Exploring machine learning methods in general and non-parametric methods in particular is now feasible and is increasingly drawing the attention of many researchers (Zhang et al., 2016; Lary et al., 2016).

More specifically, analog forecasting (Lorenz, 1969) which is among the earliest statistical methods explored in geoscience benefits from recent advances in data science. In short, analog forecasting is based on the assumption that the future state of a system can be predicted throughout the successors of past (or simulated) similar situations (called analogs). The amount of currently available remote sensing and simulation data offers analog methods a great opportunity to catch up their early promises. Several recent works involving applications of analog forecasting methods in geoscience fields contribute in the revival of these methods, recent applications comprise the prediction of soil moisture anomalies (McDermott and Wikle, 2015), the prediction of sea-ice anomalies (Comeau et al., 2017), rainfall nowcasting (Atencia

and Zawadzki, 2015), stochastic weather generators (Yiou, 2014), etc. One may also cite methodological developments such as dynamically-adapted kernels (Zhao and Giannakis, 2014a) and novel parameter estimation schemes (Horton et al., 2017). Importantly, analog strategies have recently been extended to address data assimilation issues within the so-called *analog data assimilation* (AnDA) (Lguensat et al., 2017), where the dynamical model is stated as an analog forecasting model and combined to state-of-the-art stochastic assimilation procedures such as Ensemble Kalman filters. The recent applications to high-dimensional fields in Fablet et al. (2017) provide the methodological background for this study.

Producing time-continuous and gridded maps of Sea Surface Height (SSH) is a major challenge in ocean remote sensing with important consequences on several scientific fields from weather and climate forecasting to operational needs for fisheries management and marine operations (*e.g.* Hardman-Mountford et al. (2003)). The reference gridded SSH product commonly used in the literature is distributed by the Copernicus Marine and Environment Monitoring Service (CMEMS) (formerly distributed by AVISO). This product relies on the interpolation of irregularly-spaced along-track data using an Optimal Interpolation (OI) method (Le Traon et al., 1998; Bretherton et al., 1976). While OI is relevant for the retrieval of horizontal scales of SSH fields greater than $\approx 100km$, its Gaussian assumptions cause the small scales of the SSH fields to be smoothed. This limitation makes it impossible to resolve finer-scale processes (typically from a few tens of kilometers to $\approx 100km$) which may be revealed by along-track altimetric data. This has led to a variety of research studies to improve the reconstruction of the altimetric

fields. One may cite both methodological alternatives to OI, for instance locally-adapted convolutional models (Fablet et al., 2016) and variational assimilation schemes using model-driven dynamical priors (Ubelmann et al., 2014), as well as studies exploring the synergy between different sea surface tracers, especially the synergy between SSH and SST (Sea Surface Temperature) fields and Surface Quasi-Geostrophic dynamics (Fablet et al., 2016; Klein et al., 2009; Isern-Fontanet et al., 2006, 2014; Turiel et al., 2009b,a).

In this work, we build upon our recent advances in analog data assimilation and its application to high-dimensional fields (Lguensat et al., 2017; Fablet et al., 2017). We develop an analog data assimilation model for the reconstruction of SLA fields from along-track altimeter data. It relies on a patch-based and EOF-constrained representation of the SLA fields. Using OFES numerical simulations (Masumoto et al., 2004; Sasaki et al., 2008), we design an Observation System Simulation Experiment (OSSE) for a case-study in the South China sea using real along-track sampling patterns of spaceborne altimeters. Using the resulting groundtruthed dataset, we perform a qualitative and quantitative evaluation of the proposed scheme, including comparisons to state-of-the-art schemes.

The remainder of the paper is organized as follows: Section 2 presents the different datasets used in this paper to design an OSSE, Section 3 gives insights on the classical methods used for mapping SLA from along track data, Section 4 introduces the proposed analog data assimilation model. Experimental results for the considered OSSE are shown in Section 5, and Section 6 further discuss the key aspects of this work.

2. Data: OFES (OGCM for the Earth Simulator)

An Observation System Simulation Experiment (OSSE) based on numerical simulations is considered to assess the relevance of the proposed analog assimilation framework. Our OSSE uses these numerical simulations as a groundtruthed dataset from which simulated along-track data are produced. We describe further the data preparation setup in the following sections.

2.1. Model simulation data

The Ocean General Circulation Model (OGCM) for the Earth Simulator (OFES) is considered in this study as the true state of the ocean. The simulation data is described in Masumoto et al. (2004); Sasaki et al. (2008). The coverage of the model is 75°S-75°N with a horizontal resolution of 1/10°. 34 years (1979-2012) of 3-daily simulation of SSH maps are considered, we proceed to a subtraction of a temporal mean to obtain SLA fields. In this study, our region of interest is located in the South China Sea (105°E to 117°E, 5°N to 25°N). This dataset is split into a training dataset corresponding to the first 33 years (4017 SLA maps) and a test dataset corresponding to the last year of the time series (122 SLA maps).

2.2. Along track data

We consider a realistic situation with a high rate of along track data. More precisely we use along-track data positions registered in 2014 where 4 satellites (Jason2, Cryosat2, Saral/AltiKa, HY-2A) were operating. Data is distributed by Copernicus Marine and Environment Monitoring Service (CMEMS).

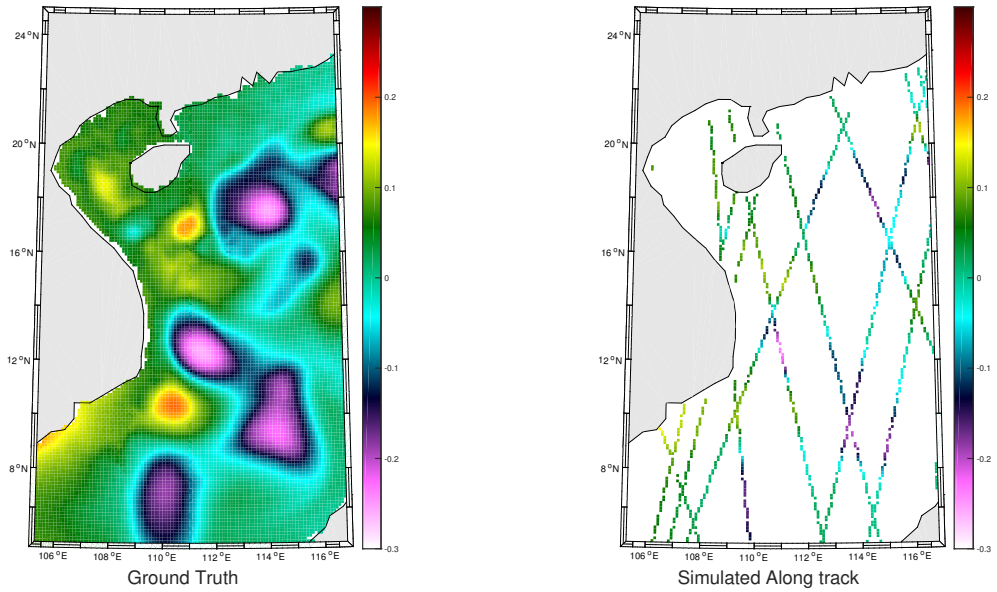


Figure 1: An example of a ground-truth SLA field in the considered region and its associated simulated pseudo-along track.

From the reference 3-daily SLA dataset and real along-track data positions, we generate simulated along-track data from the sampling of a reference SLA field: more precisely, for a given along-track point, we sample the closest position of the $1/10°$ regular model grid at the closest time step of the 3-daily model time series. As we consider a 3-daily assimilation time step (see Section 2.1 for details), we create a 3-daily pseudo-observation field, to be fed directly to the assimilation model. As sketched in Figure 2, for a given time t, we combine all along-track positions for times $t-1$,t and $t+1$ to create an along-track pseudo-observation field at time t. We denote by $s3dAT$ the simulated 3-daily time series of along-track pseudo-observation fields. An example of these fields is given in Figure 1.

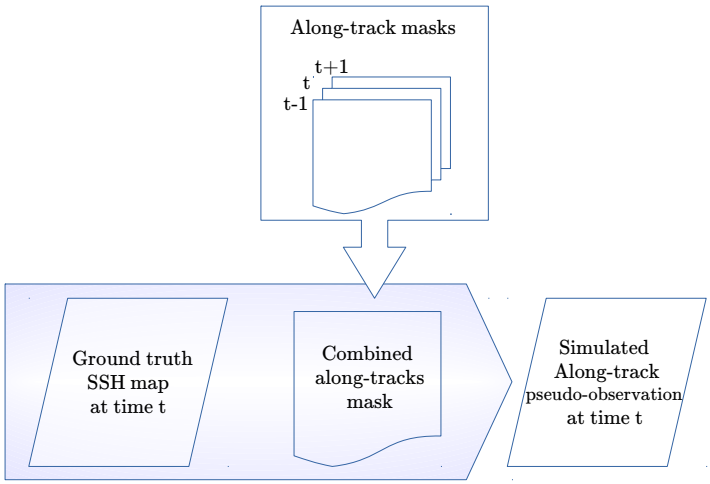


Figure 2: Sketch of the creation of simulated along-track data at a given time t

3. Problem statement and related work

3.1. Data assimilation and optimal interpolation

Data assimilation consists in estimating the true state of a physical variable $\mathbf{x}(t)$ at a specific time t, by combining i) equations governing the dynamics of the variable, ii) available observations $\mathbf{y}(1, .., T)$ measuring the variable and iii) a background or first guess on its initial state $\mathbf{x}^b$. The estimated state is generally called the analyzed state and noted by $\mathbf{x}^a$. Data assimilation is a typical example of inverse problems, and similar formulations are known to the statistical signal processing community through optimal control and estimation theory (Bocquet et al., 2010). We adopt here the unified notation of Ide et al. (1997) and formulate the problem as a stochastic system in the following:

$$\begin{cases} \mathbf{x}(t) = \mathcal{M}(\mathbf{x}(t-1)) + \eta(t), & (1) \\ \mathbf{y}(t) = \mathcal{H}(\mathbf{x}(t)) + \epsilon(t). & (2) \end{cases}$$

Equation 1 represents the dynamical model governing the evolution of state $\mathbf{x}$ through time, while η is a Gaussian centered noise of covariance $\mathbf{Q}$ that models the process error. Equation 2 explains the relationship between the observation $\mathbf{y}(t)$ and the state to be estimated $\mathbf{x}(t)$ through the operator $\mathcal{H}$. The uncertainty of the observation model is represented by the ϵ error, considered here to be Gaussian centered and of covariance $\mathbf{R}$. We assume that ϵ and η are independent and that $\mathbf{Q}$ and $\mathbf{R}$ are known. Two main approaches are generally considered for the mathematical resolution of the system (1)-(2), namely, variational data assimilation and stochastic data assimilation. They differ in the way they infer the analyzed state x^a, the first is based on the minimization of a certain cost function while the latter aims to obtain an optimal a posteriori estimate. We encourage the reader to consider the book of Asch et al. (2016) for detailed insights on the various aspects and methods of data assimilation.

A popular data assimilation algorithm that is largely used in the literature to grid sea level anomalies from along-track data is called Optimal Interpolation (OI) (e.g. Le Traon et al. (1998); De Mey and Robinson (1987)), this algorithm is also the method adopted in CMEMS altimetry product. Optimal Interpolation (OI) aims at finding the Best Linear Unbiased Estimator (BLUE) of a field $\mathbf{x}$ given irregularly sampled observations $\mathbf{y}$ in space and time and a background prior $\mathbf{x}^b$. The multivariate OI equations were derived in Gandin (1966) for meteorology and numerous applications in oceanography have been reported since the early work of Bretherton et al. (1976).

Supposing that the background state $\mathbf{x}^b$ has covariance B, and the observation operator is linear $\mathcal{H} = \mathbf{H}$, the analyzed state $\mathbf{x}^a$ and the analyzed error covariance $\mathbf{P}^a$ can be calculated using the following OI set of equations:

$$\begin{cases} \mathbf{K} = \mathbf{BH}(\mathbf{R} + \mathbf{HBH}^T)^{-1} \quad \text{called the Kalman gain} & (3) \\ \mathbf{x}^a = \mathbf{x}^b + \mathbf{K}(\mathbf{y} - \mathbf{Hx}^b) & (4) \\ \mathbf{P}^a = (\mathbf{I} - \mathbf{KH})\mathbf{B} & (5) \end{cases}$$

It worths mentioning that Lorenc (1986) showed that OI is closely related to the 3D-Var variational data assimilation algorithm which obtains $\mathbf{x}^a$ by minimizing the following cost function:

$$J(x) = (\mathbf{x} - \mathbf{x}^b)^T \mathbf{B}^{-1} (\mathbf{x} - \mathbf{x}^b) + (\mathbf{y} - \mathbf{Hx})^T \mathbf{R}^{-1} (\mathbf{y} - \mathbf{Hx}) \tag{6}$$

While OI had been shown to successfully retrieve large-scale structures in the ocean ($\geq 150km$), a well-known limitation of OI is that the Gaussian-like covariance error matrices smooths out the small-scale information (*e.g.* mesoscale eddies) (Ubelmann et al., 2014). OI would then underexploit high resolution altimeter data in the context of future altimetry missions, which urges to put efforts in trying to improve the method (e.g. Escudier et al. (2013a)) or find other alternatives.

3.2. Analog data assimilation

Endorsed by the recent development in data-driven methods and data storage capacities, the Analog Data Assimilation (AnDA) was introduced as an alternative to classical model-driven data assimilation under one or more of the following situations (Lguensat et al., 2017):

- The model is inconsistent with observations
- The cost of the model integration is high computationally
- (mandatory) The availability of large datasets of past dynamics of the variables to be estimated. These datasets are hereinafter called catalogs and denoted by $\mathcal{C}$. The catalog is organized in a two-column dictionary where each state of the system is associated with its successor in time, forming a set of couples $(\mathcal{A}_i, \mathcal{S}_i)$ where $\mathcal{A}_i$ is called the *analog* and $\mathcal{S}_i$ its *successor*.

Given the considerations above, AnDA resorts to evaluating filtering, resp. smoothing, posterior likelihood, *i.e.* the distribution of the state to be estimated $\mathbf{x}(t)$ at time t, given past and current observations $\mathbf{y}(1, .., t)$, resp. given all the available observation $\mathbf{y}(1, .., T)$. This evaluation relies on the following state-space model:

$$\begin{cases} \mathbf{x}(t) = \mathcal{F}(\mathbf{x}(t-1)) + \eta(t), & (7) \\ \mathbf{y}(t) = \mathcal{H}(\mathbf{x}(t)) + \epsilon(t). & (8) \end{cases}$$

The difference between AnDA and classical data assimilation resides in the transition model equation 7. The counterpart of a model-driven operator $\mathcal{M}$ of Equation 1 is here the operator $\mathcal{F}$ which refers to the considered data-driven operator, so called, the analog forecasting operator. This operator makes use of the available catalog $\mathcal{C}$ and assumes that the state forecast can be inferred from similar situations in the past. Provided the definitions of the analogs and successors given above, the derivation of this operator resorts to characterizing the transition distribution *i.e.* $p(\mathbf{x}(t)|\mathbf{x}(t-1))$. Following Lguensat et al. (2017), a Gaussian conditional distribution is adopted:

$$\mathbf{x}(t)|\mathbf{x}(t-1) \sim \mathcal{N}(\mu_t, \Sigma_t) \tag{9}$$

where $\mathcal{N}(\mu_t, \Sigma_t)$ is a Gaussian distribution of mean μ_t and covariance Σ_t. These parameters of the Gaussian distribution are calculated using the result of a K nearest neighbors search. The K nearest neighbors (analogs) $\mathcal{A}_{k\in(1,..,K)}$ of state $\mathbf{x}(t-1)$ and their successors $\mathcal{S}_{k\in(1,..,K)}$, along with a weight associated to each pair $(\mathcal{A}_k, \mathcal{S}_k)$ are used to calculate μ_t and Σ_t, the forecast state $\mathbf{x}(t)$ is then sampled from $\mathcal{N}(\mu_t, \Sigma_t)$. The weights are defined using a Gaussian kernel $\mathcal{K}_G$.

$$\mathcal{K}_G\left(u, v\right) = \exp\left(-\frac{\|u-v\|^2}{\sigma}\right), \tag{10}$$

Scale parameter σ is locally-adapted to the median value of the K distances $\|x(t-1) - \mathcal{A}_k\|^2$ to the K analogs. Other types of kernels might be considered (e.g. Zhao and Giannakis (2014b); McDermott and Wikle (2015)), investigating kernel choice is out of the scope of this paper.

The mean and the covariance of the transition distribution might be calculated following several strategies. We consider in this work the three analog forecasting operators introduced in AnDA (Lguensat et al., 2017):

- **Locally-constant operator:** Mean μ_t and covariance Σ_t are given by the weighted mean and covariance of the K successors $\mathcal{S}_{k\in(1,..,K)}$.

- **Locally-incremental operator:** Here, the increments between the K analogs and their corresponding successors are calculated $\mathcal{S}_{k\in(1,..,K)} - \mathcal{A}_{k\in(1,..,K)}$. The weighted mean of the K increments is then added to

the $\mathbf{x}(t-1)$ to obtain μ_t. While Σ_t results in the weighted covariance of these differences.

- **Locally-linear operator:** A weighted least-square estimation of the linear regression of the state at time t given the state at time $t-1$ is performed based on the K pairs $(\mathcal{A}_k, \mathcal{S}_k)$. The parameters of the linear regression are then applied to state $\mathbf{x}(t-1)$ to obtain μ_t. Covariance Σ_t is represented by the covariance of the residuals of the fitted weighted linear regression.

The application of the AnDA framework faces the curse of dimensionality *i.e.* the search of analogs is highly affected by the dimensionality of the problem and can be irrelevant for dimensions above 20 (Lguensat et al., 2017). As proposed in Fablet et al. (2017), the extension of AnDA models to high-dimensional fields may then rely on turning the global assimilation issue into a series of lower-dimensional ones. We consider here an approach similar to Fablet et al. (2017) using a patch-based and EOF-based representation of the 2D fields, *i.e.* the 2D fields are decomposed into a set of overlapping patches, each patch being projected onto an EOF space. Analog strategies then apply to patch-level time series in the EOF space.

Overall, as detailed in the following section, the proposed analog data assimilation model for SLA fields relies on three key components: a patch-based representation of the SLA fields, the selection of a kernel to retrieve analogs and the specification of a patch-level analog forecasting operator.

4. Analog reconstruction for altimeter-derived SLA

4.1. Patch-based state-space formulation

As stated above, OI may be considered as an efficient model-based method to recover large-scale structures of SLA fields. Following Fablet et al. (2017), this suggests considering the following two-scale additive decomposition:

$$X = \bar{X} + dX + \xi \tag{11}$$

where $\bar{X}$ is the large-scale component of the SLA field, typically issued from an optimal interpolation, dX the fine-scale component of the SLA field we aim to reconstruct and ξ remaining unresolved scales.

The reconstruction of field dX involves a patch-based and EOF-based representation. It consists in regarding field dX as a set of $P \times P$ overlapping patches (*e.g.* $2^\circ \times 2^\circ$). This set of patches is referred to as $\mathcal{P}$, and we denote by $\mathcal{P}_s$ the patch centered on position s. After building a catalog $\mathcal{C}_\mathcal{P}$ of patches from the available dataset of residual fields $X - \bar{X}$ (see Section 3.2), we proceed to an EOF decomposition of each patch in the catalog. The reconstruction of field $dX(\mathcal{P}_s, t)$ at time t is then stated as the analog assimilation of the coefficients of the EOF decomposition in the EOF space given an observation series in the patch space. Formally, $dX(\mathcal{P}_s, t)$ decomposes as a linear combination of a number N_E of EOF basis functions:

$$dX(\mathcal{P}_s, t) = \sum_{k=1}^{N_E} \alpha_k(s,t) EOF_k \tag{12}$$

with EOF_k the k^{th} EOF basis and $\alpha_k(s,t)$ the corresponding coefficient for patch $\mathcal{P}_s$ at time t. Let us denote by $\Phi(\mathcal{P}_s, t)$ the vector of the N_E coefficients

$\alpha_k(s,t)$. This vector represents the projection of $dX(\mathcal{P}_s,t)$ in the lower-dimensional EOF space.

4.2. Patch-based analog dynamical models

Given the considered patch-based representation of field dX, the proposed patch-based analog assimilation scheme involves a dynamical model stated in the EOF space. Formally, Equation 9 leads to the following Gaussian conditional distribution in the EOF space

$$\Phi(\mathcal{P}_s,t)|\Phi(\mathcal{P}_s,t-1) \sim \mathcal{N}(\mu(s,t),\Sigma(s,t)) \tag{13}$$

We consider the three analog forecasting operators presented in Section 3.2, namely, the locally-constant, the locally incremental and the locally-linear. The calculation of the weights associated to each analog-successor pair relies on a Gaussian kernel $\mathcal{K}_G$ (Equation 10). The search for analogs in the N_E-dimensional patch space (in practice, N_E ranges from 5 to 20) ensures a better accuracy in the retrieval of relevant analogs compared to a direct search in the high-dimensional space of state dX. It also reduces the computational complexity of the proposed scheme.

Another important extension of the current study is the possibility of exploiting auxiliary variables with the state vector Φ in the analog forecasting models. Such variables may be considered in the search for analogs as well as regression variables in locally-linear analog setting. Regarding the targeted application to the reconstruction of SSH fields and the proposed two-scale decomposition (Equation 11), two types of auxiliary variables seem to be of interest: the low-resolution component $\bar{X}$ to take into account inter-scale

relationship (Fablet et al., 2016), and Sea Surface Temperature (SST) with respect to the widely acknowledged SST-SSH synergies (Fablet et al., 2016; Klein et al., 2009; Isern-Fontanet et al., 2014). We also apply patch-level EOF-based decompositions to include both types of variables in the considered analog forecasting models (Equation 13).

4.3. Numerical resolution

Given the proposed analog assimilation model, the proposed scheme first relies on the creation of patch-level catalogs from the training dataset. This step requires the computation of a training dataset of fine scale data $dX_{training}$, this is done by subtracting a large-scale component $\bar{X}_{training}$ from the original training dataset. Here, we consider the large-scale component of training data to be the result of a global[1] EOF-based reconstruction using a number of EOF components that retains 95% of the dataset variance, which accounts for horizontal scales up to $\sim$ 100km. This global EOF-based decomposition provides a computationally-efficient means for defining large-scale component $\bar{X}$. This EOF-based decomposition step is followed by the extraction of overlapping patches for all variables of interest, namely $\bar{X}_{training}$, $dX_{training}$ and potential auxiliary variables, and the identification of the EOF basis functions from the resulting raw patch datasets. This leads to the creation of a patch-level catalog $\mathcal{C}_{\mathcal{P}}$ from the EOF-based representations of each patch.

Given the patch-level catalog, the algorithm applied for the mapping SLA fields from along-track data, referred to PB-AnDA (for Patch-Based AnDA),

[1]By global, we mean here an EOF decomposition over the entire case study region, by contrast to the patch-level decomposition considered in the analog assimilation setting.

involves the following steps:

- the computation of the large-scale component $\bar{X}$, here, we consider the result of optimal interpolation (OI) projected onto the global EOF basis functions.

- the decomposition of the case study region into overlapping $P \times P$ patches, here, 20×20 patches

- For each patch position s, the application of an analog data assimilation scheme, namely the Analog Ensemble Kalman Smoother (AnEnKS) (Lguensat et al., 2017), for patch $\mathcal{P}_s$ of field dX. As stated in (13), the assimilation is performed in the EOF space, *i.e.* for EOF decomposition $\Phi(\mathcal{P}_s, t)$, using the operator derived from EOF-based reconstruction (12) and decomposition (11) as observation model $\mathcal{H}$ in (8) and the patch-level training catalog described in the previous section. In the analog forecasting setting, The search for analogs is restricted to patch exemplars in the catalog within a local spatial neighborhood (typically a patch-level 8-neighborhood), except for patches along the seashore for which the search for analogs is restricted to patch exemplars at the same location.

- the reconstruction of fields dX from the set of assimilated patches $\{dX(\mathcal{P}_s, \cdot)\}_s$. This relies on a spatial averaging over overlapping patches (here, a 5-pixel overlapping in both directions). In practice, we do not apply the patch-level assimilation to all grid positions. Consequently, the spatial averaging may result in blocky artifacts. We then apply

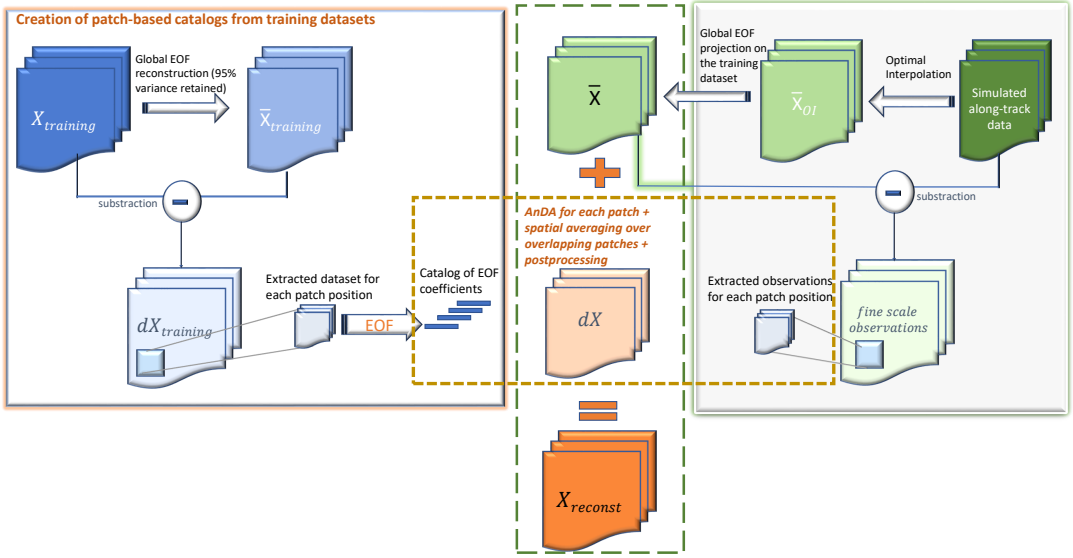


Figure 3: Sketch of the proposed patch-based Analog Data Assimilation (PB-AnDA). The left block details the construction of the patch-based catalogs from the training dataset. The right block illustrates the process of obtaining the large-scale component of the SLA reconstructed field. The orange dashed rectangle represents the application of the AnDA using the catalog and the fine-scale observations. Finally, the green dashed rectangle shows the final addition operation that yields the reconstructed SLA field.

a patchwise EOF-based decomposition-reconstruction with a smaller patch-size (here, 17×17 patches) to remove these blocky artifacts.

- the reconstruction of fields X as $\bar{X} + dX$.

5. Results

We evaluate the proposed PB-AnDA approach using the OSSE presented in Section 2. We perform a qualitative and quantitative comparison to state-of-the-art approaches. We first describe the parameter setting used for the

PB-AnDA as well as benchmarked models, namely OI, an EOF-based approach (Ping et al., 2016) and a direct application of AnDA at the region level. We then report numerical experiments for noise-free and noisy observation data as well the relevance of auxiliary variables in the proposed PB-AnDA scheme.

5.1. Experimental setting

We detail below the parameter setting of the models evaluated in the reported experiments, including the proposed PB-AnDA scheme:

- *PB-AnDA*: We consider 20×20 patches with 15-dimensional EOF decompositions ($N_E = 15$), which typically accounts for 99% of the data variance for the considered dataset. The postprocessing step exploits 17×17 patches and a 15-dimensional EOF decomposition. Regarding the parametrization of the AnEnKS procedure, we experimentally cross-validated the number of nearest neighbors K to 50, the number of ensemble members $n_{ensemble}$ to 100 and the observation covariance error in Equation 8 to $\mathbf{R} = 0.001$.

- *Optimal Interpolation*: We apply an Optimal Interpolation to the processed along-track data. It provides the low-resolution component for the proposed PB-AnDA model and a model-driven reference for evaluation purposes. The background field is a null field. We use a Gaussian covariance model with a spatial correlation length of 100km and a temporal correlation length of 15 days ($\pm$ 5 timesteps since our data is 3-daily). These choices result from a cross-validation experiment.

- *VE-DINEOF*: We apply a second state-of-the-art interpolation scheme using a data-driven strategy solely based on EOF decompositions, namely VE-DINEOF (Ping et al., 2016). We implement a patch-based version of VE-DINEOF to make it comparable to the proposed PB-AnDA setting. Given the same EOF decomposition as in PB-AnDA, the patch-level VE-DINEOF iterates patchwise EOF projection-reconstruction of the detail field dX. This scheme is initialized from the along-track pseudo-observation field for along-track data positions and $\bar{X}$ for missing data positions. After each projection-reconstruction, we only update missing data areas. We run this iterative process until convergence.

- *G-AnDA*: With a view to evaluating the relevance of the patch-based decomposition, we also apply AnDA at the region scale, referred to as G-AnDA. It relies on an EOF-based decomposition of the detail field dX. We use 150 EOF components, which accounts for more than 99% of the total variance of the SSH dataset. From cross-validation experiments, the associated AnEnKS procedure relies on a locally-linear analog forecasting model with $K = 500$ analogs, $n_{ensemble} = 100$ ensemble members and an observation covariance error in Equation 8 set to $\mathbf{R} = 0.001$

The patch-based experiments were run on Teralab infrastructure using a multi-core virtual machine (30 CPUs, 64G of RAM). We used the Python toolbox for patch-based analog data assimilation (Fablet et al., 2017) (available at github.com/rfablet/PB_ANDA). Optimal Interpolation was implemented on Matlab using Escudier et al. (2013b). Throughout the exper-

iments, two metrics are used to assess the performance of the considered interpolation methods: i) daily and mean Root Mean Square Error (RMSE) series between the reconstructed SLA fields X and the groundtruthed ones, ii) daily and mean correlation coefficient between the fine-scale component dX of the reconstructed SLA fields and of the groundtruthed ones.

5.2. SLA reconstruction from noise-free along-track data

We first perform an idealized noise-free experiment, where the along-track observations are noise-free. The observation covariance error in Equation 8 takes the value $\mathbf{R} = 0.001$. The interpolation performances for this experiment are illustrated in Table 1. Our PB-AnDA algorithm significantly outperforms OI. More specifically, the locally-linear PB-AnDA results in the best reconstruction among the competing methods. We suggest that this improvement comes from the reconstruction of fine-scale features learned from the archived model simulation data. Figure 4a reports interpolated SSH fields and their gradient fields which further confirm our intuition. PB-AnDA interpolation shows an enhancement of the gradients and comes out with some fine-scale eddies that were smoothed out in OI and VE-DINEOF. This is also confirmed by the Fourier power spectrum of the interpolated SLA fields in Figure 4b.

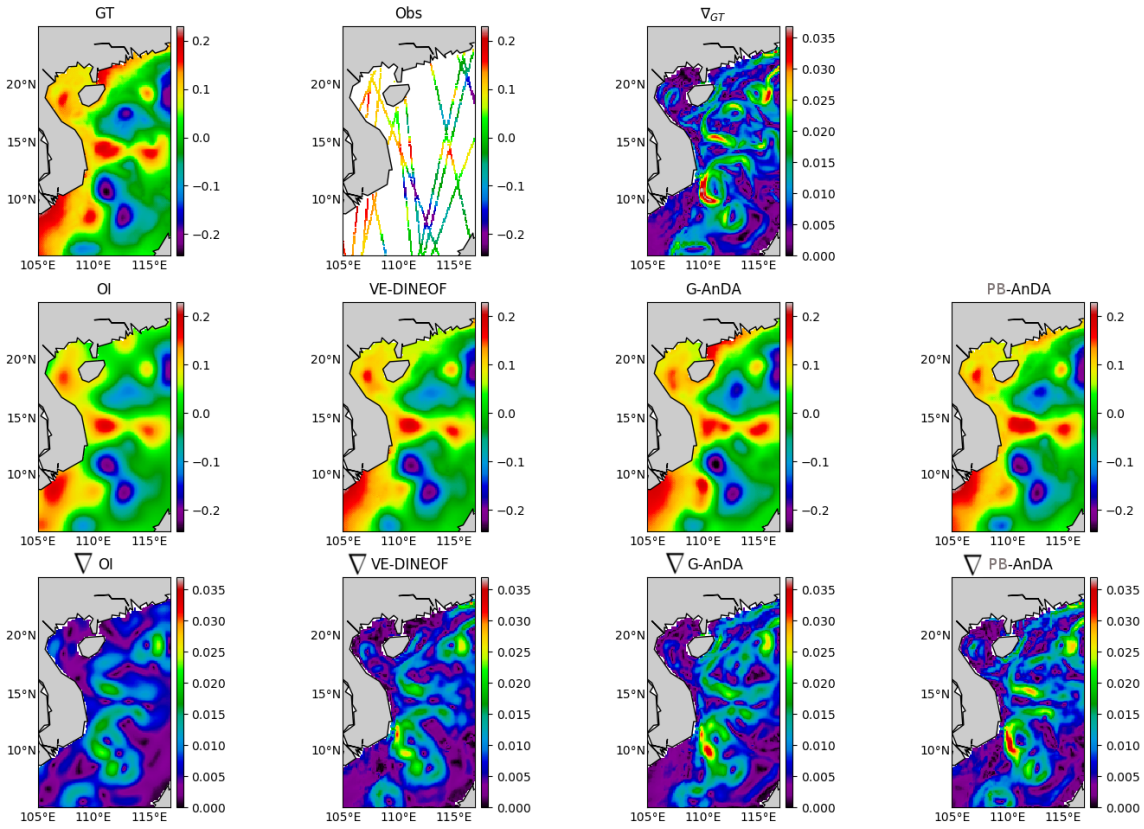
Table 1: SLA Interpolation performance for a noise-free experiment: Root Mean Square Error (RMSE) and correlation statistics for OI, VE-DINEOF, G-AnDA and PB-AnDA w.r.t. the groundtruthed SLA fields. See Section 5.1 for the corresponding parameter settings.

Criterion		RMSE	Correlation
OI		0.026 ± 0.007	0.81 ± 0.08
VE-DINEOF		0.023 ± 0.007	0.85 ± 0.07
G-AnDA		0.020 ± 0.006	0.89 ± 0.04
PB-AnDA	Locally-constant	0.014 ± 0.005	0.95 ± 0.03
	Locally-Increment	0.014 ± 0.005	0.95 ± 0.03
	Locally-Linear	$\mathbf{0.013 \pm 0.005}$	$\mathbf{0.96 \pm 0.02}$

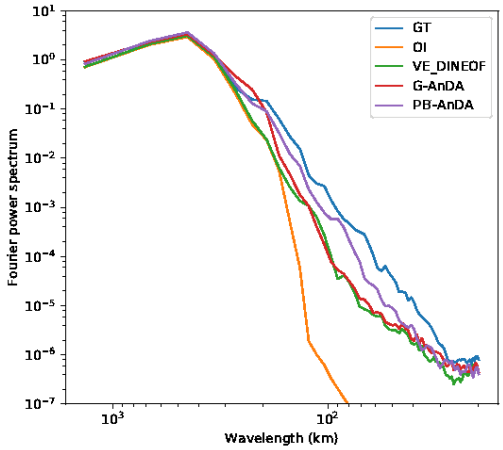
5.3. SLA reconstruction from noisy along-track data

We also evaluated the proposed approach for noisy along-track data. Here, we run two experiments with an additive zero-mean Gaussian noise applied to the simulated along-track data. We consider a noise covariance of $\mathbf{R} = 0.01$ (Experiment A) and of $\mathbf{R} = 0.03$ (Experiment B) which is more close to the instrumental error of conventional altimeters. Given the resulting noisy along-track dataset, we apply the same methods as for the noise-free case study.

We run PB-AnDA using different values for $\mathbf{R}$. For Experiment A, Table 2 shows that the minimum is reached using the true value of the error $\mathbf{R} = 0.01$. While for Experiment B, Table 3 shows that the minimum is counter-



(a)



(b)

Figure 4: Reconstructed SLA fields using noise-free along-track observation using OI, DINEOF, G-AnDA, PB-AnDA on the 54^{th} day (February 24^{th} 2012): from left to right, the first row shows the ground truth field, the simulated available along-tracks for that day, the ground truth gradient field. The second and third rows show each of the reconstruction and their corresponding gradient fields, from left to right, OI, VE-DINEOF, G-ANDA and PB-AnDA. The Fourier power spectrum of the competing methods is also included

intuitively reached again using value of the error $\mathbf{R} = 0.01$.

Our algorithm is then compared with the results of the application of the competing algorithms considered in this work. Results are shown in Table 4. PB-AnDA still outperforms OI in terms of RMSE and correlation statistics in both experiments. The locally-linear version of PB-AnDA depicts the best reconstruction performance. We report an example of the reconstruction in Figure 5. Similarly to the noise-free case study, PB-AnDA better recovers finer-scale structures in Fig.5.a compared with OI, VE-DINEOF and G-AnDA. In Fig.5.b, PB-AnDA also better reconstructs a larger-scale North-East structure, poorly sampled by along-track data and hence poorly interpolated by OI.

Table 2: Impact of variance of observation error R in AnDA interpolation performance using noisy along-track data (R=0.01): RMSE of AnDA interpolation for different values of parameter R. For the same dataset, OI RMSE is **0.039**.

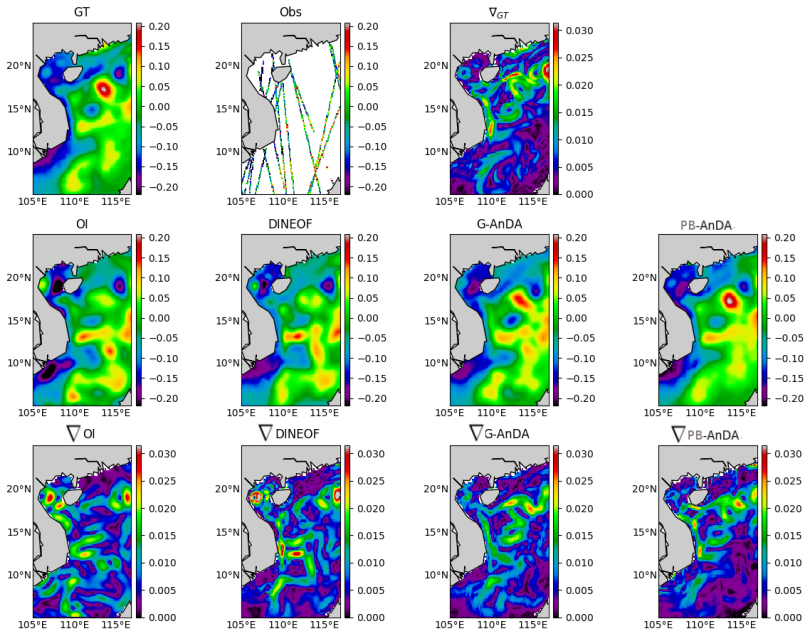
R	0.1	0.05	0.03	**0.01**	0.005	0.001	0.0001
$rmse_{PB-AnDA}$	0.035	0.030	0.028	**0.025**	0.025	0.029	0.044

Table 3: Impact of variance of observation error R in AnDA interpolation performance using noisy along-track data (R=0.03): RMSE of AnDA interpolation for different values of parameter R. For the same dataset, OI RMSE is **0.066**.

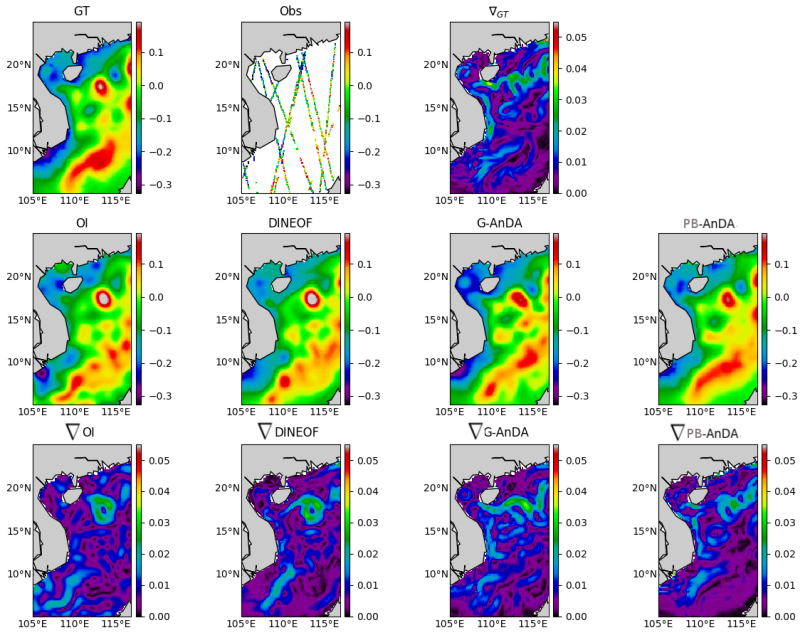
R	0.1	0.05	0.03	**0.01**	0.005	0.001	0.0001
$rmse_{PB-AnDA}$	0.038	0.036	0.035	**0.0349**	0.037	0.046	0.076

Table 4: SLA Interpolation performance for noisy along-track data: Root Mean Square Error (RMSE) and correlation statistics for OI, VE-DINEOF, G-AnDA and PB-AnDA w.r.t. the groundtruthed SLA fields. See Section 5.1 for the corresponding parameter settings.

	Criterion		RMSE	Correlation
R=0.01		OI	0.039 ± 0.005	0.64 ± 0.09
		VE-DINEOF	0.035 ± 0.005	0.68 ± 0.09
		G-AnDA	0.030 ± 0.005	0.78 ± 0.06
	PB-AnDA	Locally constant	0.026 ± 0.005	0.82 ± 0.05
		Increment	0.028 ± 0.005	0.81 ± 0.05
		Local Linear	**0.0245 ± 0.005**	**0.83 ± 0.05**
R=0.03		OI	0.066 ± 0.006	0.41 ± 0.09
		VE-DINEOF	0.060 ± 0.006	0.45 ± 0.09
		G-AnDA	0.039 ± 0.006	0.67 ± 0.09
	PB-AnDA	Locally constant	0.035 ± 0.006	0.688 ± 0.064
		Increment	0.036 ± 0.006	0.656 ± 0.07
		Local Linear	**0.032 ± 0.006**	**0.708 ± 0.063**



(a)



(b)

Figure 5: Reconstruction of SLA fields from noisy along-track data using OI, DINEOF, G-AnDA & PB-AnDA on day 225^{th} (a) & 228^{th} (b)

5.4. PB-AnDA models with auxiliary variables

We further explore the flexibility of the analog setting to the use of additional geophysical variable information as explained in Section 4.2. Intuitively, we expect SLA fields to involve inter-scale dependencies as well as synergies with other tracers. The use of auxiliary variables provide the means for evaluating such dependencies and their potential impact on reconstruction performance. We consider two auxiliary variables that are used in the locally-linear analog forecasting model (7): i) to account for the relationship between the large-scale and fine-scale component, we may consider variable $\bar{X}$, ii) considering potential SST-SSH synergies, we consider SST fields. Overall, we consider four parameterization of the regression variables used in PB-AnDA: the sole use of dX (PB-AnDA-dX); the joint use of dX and SST fields (PB-AnDA-dX+SST); the joint use of dX and $\bar{X}$ (PB-AnDA-dX+$\bar{X}$), the joint use of dX and the groudntruthed version of $\bar{X}$ denoted by $\bar{X}^{GT}$, (PB-AnDA-dX+$\bar{X}^{GT}$). The later provides a lower-bound for the reconstruction performance, assuming the low-resolution component is perfectly estimated.

We report mean RMSE and correlation statistics for these four PB-AnDA parameterizations in Table 5 for the noisy case-study. Considering PB-AnDA-dX as reference, these results show a very slight improvement when complementing dX with SST information. Though limited, we report a greater improvement when adding the low-resolution component $\bar{X}$. Interestingly, a significantly greater improvement is obtained when adding the true low-resolution information. The mean results are in accordance with Fablet et al. (2016), which reported that large-scale SLA information was

more informative than SST to improve the reconstruction of the SLA at finer scales. Though mean statistics over one year leads to rather limited improvement, daily RMSE time series (Figure 6) reveal that for some periods, for instance between day 130 and 150, relative improvements in terms of RMSE may reach 10% with the additional information brought by the large-scale component. In this respect, it may noted that PB-AnDA-dX+$\bar{X}$ always perform better than PB-AnDA-dX.

Table 5: PB-AnDA reconstruction performance using noisy along-track data for different choices of the regression variables in the locally-linear analog forecasting model: PB-AnDA-dX using solely dX, PB-AnDA-dX+SST using both dX and SST, PB-AnDA-$dX+\bar{X}$ using both dX and $\bar{X}$, and PB-AnDA-$dX+\bar{X}^{GT}$ using dX and the true large-scale component $\bar{X}^{GT}$.

	PB-AnDA model	RMSE	Correlation
R=0.01	PB-AnDA-dX	0.025 ± 0.005	0.83 ± 0.05
	PB-AnDA-dX+SST	0.024 ± 0.005	0.83 ± 0.05
	PB-AnDA-$dX + \bar{X}$	0.023 ± 0.005	0.84 ± 0.05
	PB-AnDA-$dX + \bar{X}^{GT}$	0.021 ± 0.004	0.87 ± 0.04
R=0.03	PB-AnDA-dX	0.032 ± 0.006	0.708 ± 0.06
	PB-AnDA-dX+SST	0.031 ± 0.006	0.710 ± 0.06
	PB-AnDA-$dX + \bar{X}$	0.029 ± 0.006	0.717 ± 0.06
	PB-AnDA-$dX + \bar{X}^{GT}$	0.026 ± 0.005	0.730 ± 0.05

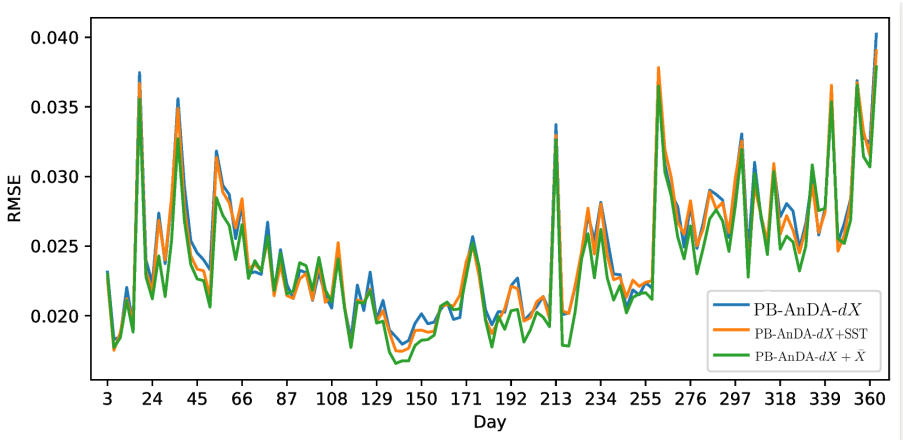
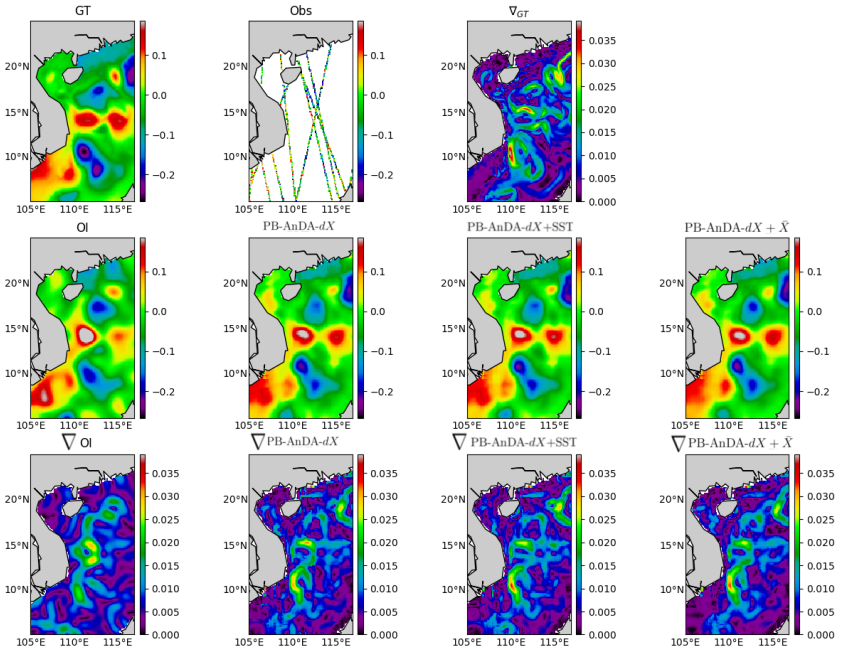
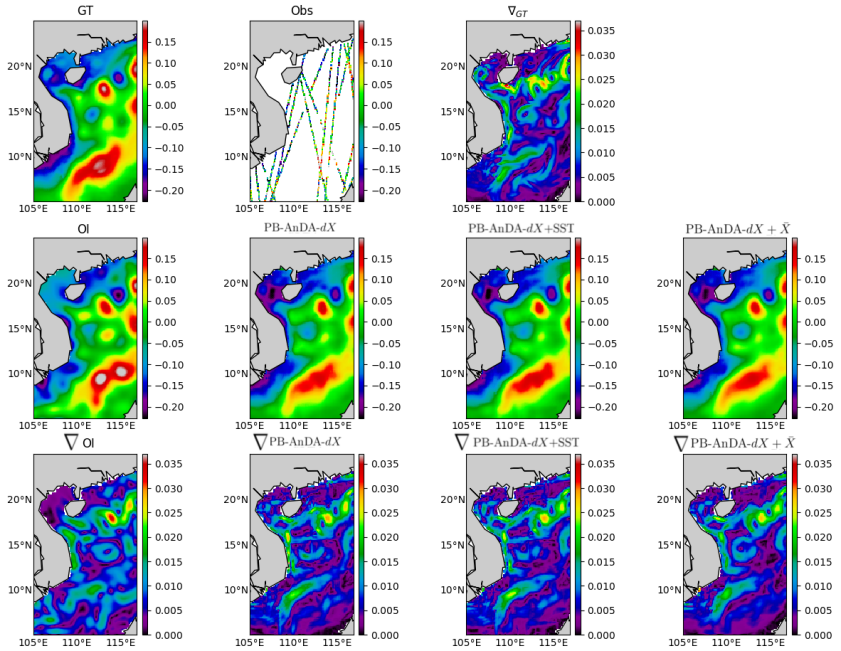


Figure 6: Daily RMSE time series of PB-AnDA SLA reconstructions using noisy along-track data for different choices of the regression variables in the locally-linear analog forecasting model: PB-AnDA-dX (light blue), PB-AnDA-dX+SST (orange) and PB-AnDA-$dX + \bar{X}$ (green).



(a)



(b)

Figure 7: (Noisy observation) Reconstruction of SLA fields using PB-AnDA with different multivariate regression models on day 57^{th} & 237^{th} (b)

6. Discussion and conclusion

This work sheds light on the opportunities that data science methods are offering to improve altimetry in the era of "Big Data". Assuming the availability of high-resolution numerical simulations, we show that Analog Data Assimilation (AnDA) can outperform the Optimal Interpolation method and retrieve smoothed out structures resulting from the sole use of OI both with idealized noise-free and more realistic noisy observations for the considered case study. Importantly, the reported experiments point out the relevance for combining OI for larger scales (above 100km) whereas the proposed patch-based analog setting successfully applies to the finer-scale range below 100km. This is in agreement with the recent application of the analog data assimilation to the reconstruction of cloud-free SST fields (Fablet et al., 2017). We also demonstrate that AnDA can embed complementary variables in a simple manner through the regression variables used in the locally-linear analog forecasting operator. In agreement with our recent analysis (Fablet et al., 2016), we demonstrate that the additional use of local SST and large-scale SLA information may further improve the reconstruction performance for fine-scale structures.

Analog data assimilation can be regarded as a means to fuse ocean models and satellite-derived data. We regard this study as a proof-of-concept, which opens research avenues as well as new directions for operational oceanography. Our results advocate for complementary experiments at the global scale or in different ocean regions for a variety of dynamical situations with a view to further evaluating the relevance of the proposed analog assimilation framework. Such experiments should evaluate the sensitivity of the assimilation

with respect to the size of the catalog. The scaling up to the global ocean also suggests investigating computationally-efficient implementation of the analog data assimilation. In this respect, the proposed patch-based framework intrinsically ensures high parallelization performance. From a methodological point of view, a relative weakness of the analog forecasting models (9) may be their low physical interpretation compared with physically-derived priors (Ubelmann et al., 2014). The combination of such physically-derived parameterizations to data-driven strategies appear as a promising research direction.

Beyond along-track altimeter data as considered in this study, future missions such as SWOT (NASA/CNES) promise an unprecedented coverage around the globe. More specifically, the large swath is expected to provide a large number of data, urging for the inspection of the potential improvements that this new mission will bring compared to classical along-track data. In the context of analog data assimilation, the interest of SWOT data may be two-fold. First, regarding observation model (8), SWOT mission will both significantly increase the number of available observation data and enable the definition of more complex observation models exploiting for instance velocity-based or vorticity-based criterion. Second, SWOT data might also be used to build representative patch-level catalogs of exemplars. Future work should investigate these two directions using simulated SWOT test-beds (Gaultier et al., 2015).

Acknowledgments

This work was supported by ANR (Agence Nationale de la Recherche, grant ANR-13-MONU-0014), Labex Cominlabs (grant SEACS) and TeraLab (grant TIAMSEA).

Asch, M., Bocquet, M., Nodet, M., 2016. Data assimilation: methods, algorithms, and applications. Fundamentals of Algorithms. SIAM.
URL https://hal.inria.fr/hal-01402885

Atencia, A., Zawadzki, I., 2015. A comparison of two techniques for generating nowcasting ensembles. part ii: Analogs selection and comparison of techniques. Monthly Weather Review 143 (7), 2890–2908.

Bocquet, M., Pires, C. A., Wu, L., 2010. Beyond gaussian statistical modeling in geophysical data assimilation. Monthly Weather Review 138 (8), 2997–3023.

Bretherton, F. P., Davis, R. E., Fandry, C., 1976. A technique for objective analysis and design of oceanographic experiments applied to mode-73. In: Deep Sea Research and Oceanographic Abstracts. Vol. 23. Elsevier, pp. 559–582.

Comeau, D., Giannakis, D., Zhao, Z., Majda, A. J., 2017. Predicting regional and pan-arctic sea ice anomalies with kernel analog forecasting. arXiv preprint arXiv:1705.05228.

De Mey, P., Robinson, A. R., 1987. Assimilation of altimeter eddy fields in a limited-area quasi-geostrophic model. Journal of physical oceanography 17 (12), 2280–2293.

505 Escudier, R., Bouffard, J., Pascual, A., Poulain, P.-M., Pujol, M.-I., 2013a.
506 Improvement of coastal and mesoscale observation from space: Applica-
507 tion to the northwestern mediterranean sea. Geophysical Research Letters
508 40 (10), 2148–2153.

509 Escudier, R., Bouffard, J., Pascual, A., Poulain, P.-M., Pujol, M.-I.,
510 2013b. Improvement of coastal and mesoscale observation from space:
511 Application to the northwestern Mediterranean Sea. Geophysical Research
512 Letters 40 (10), 2148–2153.
513 URL http://onlinelibrary.wiley.com/doi/10.1002/grl.50324/full

514 Fablet, R., Verron, J., Mourre, B., Chapron, B., Pascual, A., Oct. 2016. Im-
515 proving mesoscale altimetric data from a multi-tracer convolutional pro-
516 cessing of standard satellite-derived products, working paper or preprint.
517 URL https://hal.archives-ouvertes.fr/hal-01365761

518 Fablet, R., Viet, P. H., Lguensat, R., 2017. Data-driven Models for the
519 Spatio-Temporal Interpolation of satellite-derived SST Fields. IEEE Trans-
520 actions on Computational Imaging.

521 Gandin, L., 1966. Objective analysis of meteorological fields. by L. S. Gandin.
522 translated from the russian. jerusalem (israel program for scientific trans-
523 lations), 1965. pp. vi, 242: 53 figures; 28 tables. 4 1s. 0d. Quarterly Journal
524 of the Royal Meteorological Society 92 (393), 447–447.
525 URL http://dx.doi.org/10.1002/qj.49709239320

526 Gaultier, L., Ubelmann, C., Fu, L.-L., Nov. 2015. The Challenge of Using

Future SWOT Data for Oceanic Field Reconstruction. Journal of Atmospheric and Oceanic Technology 33 (1), 119–126.

Hardman-Mountford, N., Richardson, A., Boyer, D., Kreiner, A., Boyer, H., 2003. Relating sardine recruitment in the northern benguela to satellite-derived sea surface height using a neural network pattern recognition approach. Progress in Oceanography 59 (2), 241 – 255, eNVIFISH: Investigating environmental causes of pelagic fisheries variability in the SE Atlantic.
URL http://www.sciencedirect.com/science/article/pii/S0079661103001721

Horton, P., Jaboyedoff, M., Obled, C., 2017. Global optimization of an analog method by means of genetic algorithms. Monthly Weather Review 145 (4), 1275–1294.

Ide, K., Courtier, P., Ghil, M., Lorenc, A., 1997. Unified notation for data assimilation: operational, sequential and variational. Practice 75 (1B), 181–189.

Isern-Fontanet, J., Chapron, B., Lapeyre, G., Klein, P., 2006. Potential use of microwave sea surface temperatures for the estimation of ocean currents. GEOPHYSICAL RESEARCH LETTERS 33, l24608.
URL doi:10.1029/2006GL027801

Isern-Fontanet, J., Shinde, M., Andersson, C., Mar. 2014. On the Transfer Function between Surface Fields and the Geostrophic Stream Function in the Mediterranean Sea. Journal of Physical Oceanography 44 (5),

549 1406–1423.
550 URL http://journals.ametsoc.org/doi/abs/10.1175/JPO-D-13-0186.1

551 Klein, P., Isern-Fontanet, J., Lapeyre, G., Roullet, G., Danioux, E.,
552 Chapron, B., Le Gentil, S., Sasaki, H., Jun. 2009. Diagnosis of vertical
553 velocities in the upper ocean from high resolution sea surface height.
554 Geophysical Research Letters 36 (12), L12603.
555 URL http://onlinelibrary.wiley.com/doi/10.1029/2009GL038359/abstract

556 Lary, D. J., Alavi, A. H., Gandomi, A. H., Walker, A. L., 2016. Machine
557 learning in geosciences and remote sensing. Geoscience Frontiers 7 (1),
558 3–10.

559 Le Traon, P., Nadal, F., Ducet, N., 1998. An improved mapping method of
560 multisatellite altimeter data. Journal of Atmospheric and Oceanic Tech-
561 nology 15 (2), 522–534.

562 Lguensat, R., Tandeo, P., Ailliot, P., Pulido, M., Fablet, R., 2017. The analog
563 data assimilation. Monthly Weather Review 0 (0), null.
564 URL https://doi.org/10.1175/MWR-D-16-0441.1

565 Lorenc, A. C., 1986. Analysis methods for numerical weather prediction.
566 Quarterly Journal of the Royal Meteorological Society 112 (474), 1177–
567 1194.

568 Lorenz, E. N., 1969. Atmospheric predictability as revealed by naturally oc-
569 curring analogues. Journal of the Atmospheric sciences 26 (4), 636–646.

570 Masumoto, Y., Sasaki, H., Kagimoto, T., Komori, N., Ishida, A., Sasai, Y.,

Miyama, T., Motoi, T., Mitsudera, H., Takahashi, K., et al., 2004. A fifty-year eddy-resolving simulation of the world ocean: Preliminary outcomes of ofes (ogcm for the earth simulator). J. Earth Simulator 1, 35–56.

McDermott, P. L., Wikle, C. K., 2015. A model-based approach for analog spatio-temporal dynamic forecasting. Environmetrics, n/a–n/a.
URL http://dx.doi.org/10.1002/env.2374

Ping, B., Su, F., Meng, Y., May 2016. An Improved DINEOF Algorithm for Filling Missing Values in Spatio-Temporal Sea Surface Temperature Data. PLOS ONE 11 (5), e0155928.
URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0155928

Sasaki, H., Nonaka, M., Masumoto, Y., Sasai, Y., Uehara, H., Sakuma, H., 2008. An eddy-resolving hindcast simulation of the quasi-global ocean from 1950 to 2003 on the earth simulator.

Turiel, A., Nieves, V., Garcia-Ladona, E., Font, J., Rio, M.-H., Larnicol, G., 2009a. The multifractal structure of satellite sea surface temperature maps can be used to obtain global maps of streamlines. Ocean Science 5 (4), 447–460.
URL https://www.ocean-sci.net/5/447/2009/

Turiel, A., Sole, J., Nieves, V., Ballabrera-Poy, J., Garcia-Ladona, E., 2009b. Tracking oceanic currents by singularity analysis of Microwave Sea Surface Temperature images. Remote Sensing of Environment In Press.
URL http://www.sciencedirect.com/science/article/B6V6V-4RFD07Y-1/2/9a77ceb1b616

593 Ubelmann, C., Klein, P., Fu, L.-L., Oct. 2014. Dynamic Interpolation of Sea
594 Surface Height and Potential Applications for Future High-Resolution
595 Altimetry Mapping. Journal of Atmospheric and Oceanic Technology
596 32 (1), 177–184.
597 URL http://journals.ametsoc.org/doi/abs/10.1175/JTECH-D-14-00152.1

598 Yiou, P., 2014. Anawege: a weather generator based on analogues of atmo-
599 spheric circulation. Geoscientific Model Development 7 (2), 531–543.

600 Zhang, L., Zhang, L., Du, B., 2016. Deep learning for remote sensing data:
601 A technical tutorial on the state of the art. IEEE Geoscience and Remote
602 Sensing Magazine 4 (2), 22–40.

603 Zhao, Z., Giannakis, D., Dec. 2014a. Analog Forecasting with Dynamics-
604 Adapted Kernels. arXiv:1412.3831 [physics]ArXiv: 1412.3831.
605 URL http://arxiv.org/abs/1412.3831

606 Zhao, Z., Giannakis, D., 2014b. Analog forecasting with dynamics-adapted
607 kernels. arXiv preprint arXiv:1412.3831.