



Article

# Data-Driven Interpolation of Sea Level Anomalies Using Analog Data Assimilation

Redouane Lguensat <sup>1,\*</sup>, Phi Huynh Viet <sup>2</sup>, Miao Sun <sup>3</sup>, Ge Chen <sup>4</sup>, Tian Fenglin <sup>4</sup>, Bertrand Chapron <sup>5</sup> and Ronan Fablet <sup>2</sup>

<sup>1</sup> IGE, Université Grenoble Alpes, CNRS, IRD, Grenoble INP, 38000 Grenoble, France

<sup>2</sup> IMT Atlantique, Lab-STICC UMR CNRS 6285, UBL, 29200 Brest, France; Vietphi3892@gmail.com (P.H.V.); ronan.fablet@imt-atlantique.fr (R.F.)

<sup>3</sup> Key Laboratory of Digital Ocean, National Marine Data and Information Service, Tianjin 300171, China; miaomiao\_1987qq@126.com

<sup>4</sup> Department of Marine Information Technology, Ocean University of China, Qingdao 266100, China; gechen@ouc.edu.cn (G.C.); tianfenglin@ouc.edu.cn (T.F.)

<sup>5</sup> Laboratoire d'Océanographie Physique et Spatiale, IFREMER, 29200 Brest, France; bertrand.chapron@ifremer.fr

\* Correspondence: redouane.lguensat@univ-grenoble-alpes.fr

Received: 21 January 2019; Accepted: 4 April 2019; Published: 9 April 2019



**Abstract:** From 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 Sea Level Anomaly (SLA) fields using analog strategies. This reconstruction is stated as an analog data assimilation issue, where the analog models rely on patch-based and Empirical Orthogonal Functions (EOF)-based representations to circumvent the curse of dimensionality. We implement an Observation System Simulation Experiment (OSSE) 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 (RMSE) for scales below 100 km. We further discuss the usefulness of the proposed analog model as a means to exploit high-resolution model simulations for the processing and analysis of current and future satellite-derived altimetric data with regard to conventional interpolation schemes, especially optimal interpolation.

**Keywords:** analog data assimilation; sea level anomaly; sea surface height; interpolation; data-driven methods

## 1. Introduction

Over the last two decades, ocean remote sensing data has benefited from numerous remote earth observation missions. These satellites measured and transmitted data about several ocean properties, 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 the 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 [1,2].

More specifically, analog forecasting [3] 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 [4], the prediction of sea-ice anomalies [5], rainfall nowcasting [6], numerical weather prediction [7–9], etc. One may also cite methodological developments such as dynamically-adapted kernels [10] and novel parameter estimation schemes [11]. Importantly, analog strategies have recently been extended to address data assimilation issues within the so-called analog data assimilation (AnDA) [12,13], 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.

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., [14]). The reference gridded SSH product commonly used in the literature is distributed by Copernicus Marine 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 [15,16]. While OI is relevant for the retrieval of horizontal scales of SSH fields up to  $\approx 100$  km, the prescribed covariance priors lead to smoothing out finer-scales. Typically, horizontal scales from a few tens of kilometers to  $\approx 100$  km may be poorly resolved by OI-derived SSH fields, while they may be partially 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 [17] and variational assimilation schemes using model-driven dynamical priors [18], 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 [17,19–23].

In this work, we build upon our recent advances in analog data assimilation and its application to high-dimensional fields. While the works in [12,13] presented the AnDA framework by combining the analog forecasting method and stochastic filtering, these works have only shown applications to geophysical toy models. It was not until the work in [24] that the AnDA methodology was applied to realistic high dimensional fields, namely, Sea Surface Temperature (SST). Dealing with the curse of dimensionality is a critical challenge, in [24] we have shown that the use of patch-based representations (a patch is a term used by the image processing community to refer to smaller image parts of a given global image [25]) combined with EOF-based representations (EOF stands for Empirical Orthogonal Function, a classic dimensionality reduction technique also known as Principal Component Analysis (PCA)) leads to a computationally-efficient interpolation of missing data in SST maps outperforming classical OI-based interpolation schemes. Another development in AnDA applied to high dimensional fields was the introduction of conditional and physically-derived operators [26], where the analog forecasting operators account for the theoretical studies relating to synergies between ocean variables (e.g., SSH and SST) and those highlighting the importance of inter-scale dependencies. In this paper, we make use of these previously developed methodologies and tools and apply the AnDA to Sea Level Anomaly fields. The contribution of this work is two-fold: (i) Confronting AnDA to the reconstruction of an ocean tracer with scarce observations compared to SST (due to the nature of the available altimeters); (ii) designing an Observation System Simulation Experiment (OSSE) based on numerical simulation data to build the archived datasets used for the analog search; (iii) Reconstructing Sea Level Anomaly (SLA) by using SST or large scale SLA as auxiliary variables embedded in the analog regression techniques as shown in Section 4.

Using OFES (Ocean General Circulation Model (OGCM) for the Earth Simulation) numerical simulations [27,28], 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. Several particular mesoscale variation patterns characterizing this region were studied in the literature, we refer the reader to [29] and references therein. We also note that our method is not region specific and can be

applied to any region of interest. 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 settings are detailed in section 5 and results for the considered OSSE are shown in Section 6. Section 7 further discusses 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.

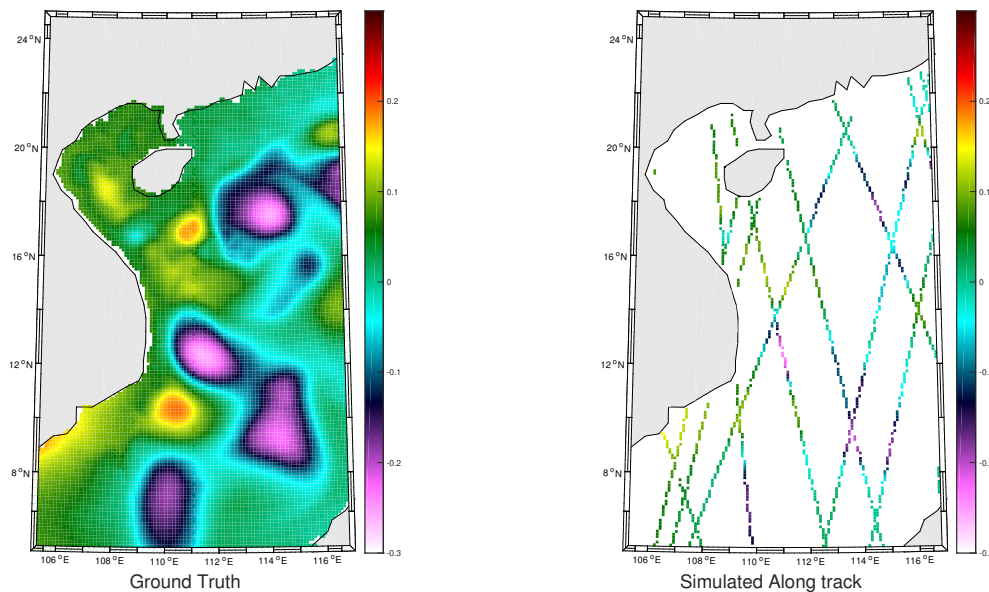
### 2.1. Model Simulation Data

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 [27,28]. 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 four satellites (Jason2, Cryosat2, Saral/AltiKa, HY-2A) were operating. Data is distributed by Copernicus Marine and Environment Monitoring Service (CMEMS).

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. 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 1.** An example of a ground-truth Sea Level Anomaly (SLA) field (meters) in the considered region and its associated simulated pseudo-along track.

### 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  $\mathbf{y}(1, \dots, T)$  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 [30]. We adopt here the unified notation of [31] 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  $\eta$  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  $\epsilon$  error, considered here to be Gaussian centered and of covariance  $\mathbf{R}$ . We assume that  $\epsilon$  and  $\eta$  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) and (2), namely, variational data assimilation and stochastic data assimilation. They differ in the way they infer the analyzed state  $\mathbf{x}^a$ , 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 [32] 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., [15,33]), 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 [34] for meteorology and numerous applications in oceanography have been reported since the early work of [16]. Supposing that the background state  $\mathbf{x}^b$  has covariance  $\mathbf{B}$ , 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:

$$\mathbf{K} = \mathbf{B}\mathbf{H}(\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T)^{-1} \quad \text{called the Kalman gain,} \quad (3)$$

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x}^b), \quad (4)$$

$$\mathbf{P}^a = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{B}. \quad (5)$$

It is worth mentioning that [35] 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:

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

An important limitation of OI is that the Gaussian-like covariance priors lead to smoothing out the small-scale information (e.g., mesoscale eddies), more specifically, this is a limitation due to the use of a static climatological  $\mathbf{B}$  matrix. For satellite-derived altimetry fields, this usually results in over-smoothing altimetry fields for structures below  $\approx 100$  km [18]. This limitation may be even more critical in the context of future high-resolution altimetry missions, which supports the development of new OI-based methods (e.g., multi-scale OI schemes as in [36]) or alternatives as addressed by our work.

### 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 [13]:

- The model is inconsistent with observations.
- The cost of the model integration is high computationally.
- (mandatory) The availability of a representative (large) dataset of the dynamics of the state 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, respectively smoothing, posterior likelihood, i.e., the distribution of the state to be estimated  $\mathbf{x}(t)$  at time  $t$ , given past and current observations  $\mathbf{y}(1, \dots, t)$ , respectively given all the available observation  $\mathbf{y}(1, \dots, T)$ . This evaluation relies on the following state-space model:

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

$$\quad (8)$$

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 [13], a Gaussian conditional distribution is adopted:

$$p(\mathbf{x}(t)|\mathbf{x}(t-1)) = \mathcal{N}(\mu(\mathbf{x}(t-1)), \Sigma(\mathbf{x}(t-1))), \quad (9)$$

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

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

Scale parameter  $\sigma$  is locally-adapted to the median value of the  $K$  distances  $\|\mathbf{x}(t-1) - \mathcal{A}_k\|^2$  to the  $K$  analogs. Other types of kernels might be considered (e.g., [4,10]), 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 [13], more details can be found in Appendix A:

- **Locally-constant operator:** Mean  $\mu(\mathbf{x}(t-1))$  and covariance  $\Sigma(\mathbf{x}(t-1))$  are given by the weighted mean and covariance of the  $K$  successors  $\mathcal{S}_{k \in (1, \dots, K)}$ .
- **Locally-incremental operator:** Here, the increments between the  $K$  analogs and their corresponding successors are calculated  $\mathcal{S}_{k \in (1, \dots, K)} - \mathcal{A}_{k \in (1, \dots, K)}$ . The weighted mean of the  $K$  increments is then added to the  $\mathbf{x}(t-1)$  to obtain  $\mu(\mathbf{x}(t-1))$ . While  $\Sigma(\mathbf{x}(t-1))$  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  $\mu(\mathbf{x}(t-1))$ . Covariance  $\Sigma(\mathbf{x}(t-1))$  is represented by the covariance of the residuals of the fitted weighted linear regression.

We may state clearly the key difference between the AnDA and reduced-rank Kalman filters and the OI method. It lies in the fact that the AnDA introduces a dynamical operator and not a prescribed space-time covariance model, and this dynamical operator is state-dependent and globally non-linear. The proposed analog forecasting operator can be seen as a state-dependent linear Gaussian operator, meaning that it is locally Gaussian and linear at each time step with a parameterization that depends on the current state, such that globally the dynamical operator is non-linear and non-Gaussian. A special case where AnDA is equivalent to OI is when all the elements of the catalog are considered as neighbors of any state vector. This case comes to assume that the dynamical operator is linear and state-independent. It is obviously of low interest due to the computational burden resulting from using all the catalog.

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 fail at finding good analogs for state vector dimensions above 20 [13]. As proposed in [24], 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 [24] using a patch-based and EOF-based representation of the two-dimensional (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 [24], this suggests considering the following two-scale additive decomposition:

$$X = \bar{X} + dX + \zeta, \tag{11}$$

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  $\zeta$  is the 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.,  $2^\circ \times 2^\circ$ ), an example of patch locations is shown in Figure 2. 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  $\alpha_k(s, t)$ :  $\Phi(\mathcal{P}_s, t) = \{\alpha_1(s, t), \dots, \alpha_{N_E}(s, t)\}$ . This vector represents the projection of  $dX(\mathcal{P}_s, t)$  in the lower-dimensional EOF space.

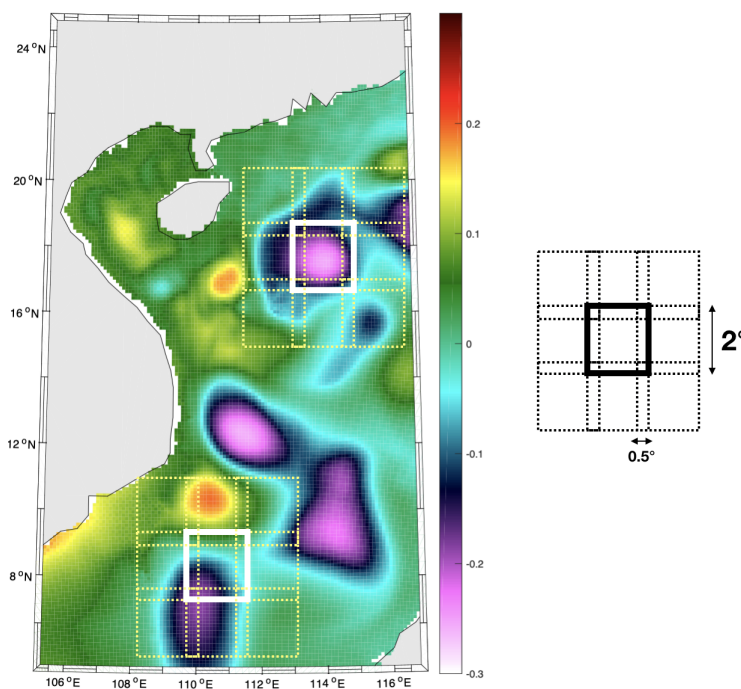


Figure 2. Two examples of patch locations and their overlapping patch neighbours.

#### 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:

$$p(\Phi(\mathcal{P}_s, t) | \Phi(\mathcal{P}_s, t - 1)) = \mathcal{N}(\mu(\Phi(\mathcal{P}_s, t - 1)), \Sigma(\Phi(\mathcal{P}_s, t - 1))), \quad (13)$$

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  $\Phi$  in the analog forecasting models. Such variables may be considered in the search for analogs as well as regression variables in a 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 [17], and Sea Surface Temperature (SST) with respect to the widely acknowledged SST-SSH synergies [17,19,21]. 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 (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.) EOF-based reconstruction using a number of EOF components that retains 95% of the dataset variance, which accounts for horizontal scales up to  $\approx 100$  km. 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  $\mathcal{C}_P$  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), is stated in Algorithm 1 and a sketch of the method is shown in Figure 3.

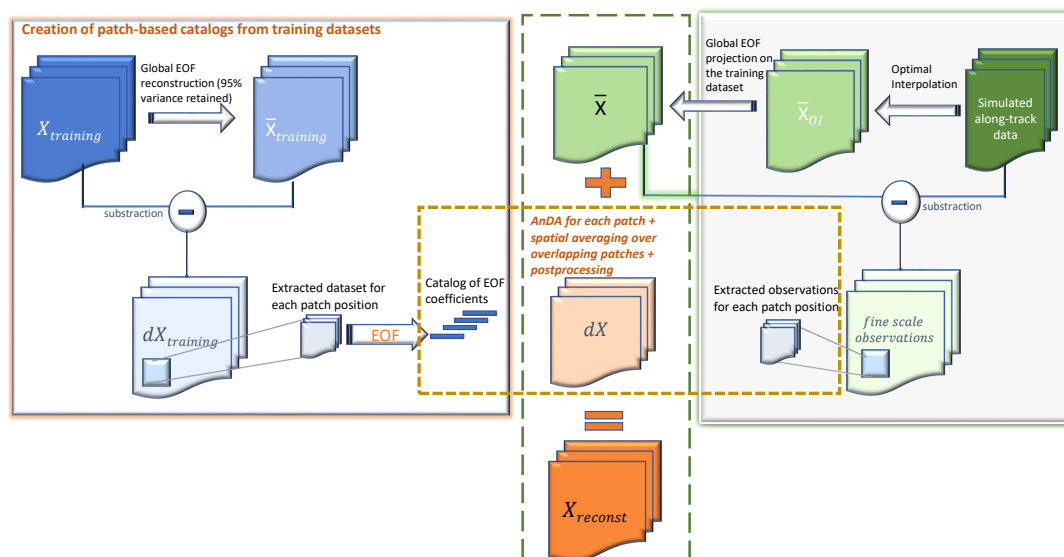


**Algorithm 1** Patch-Based AnDA

- 1: Compute the large-scale component  $\bar{X}$ , here, we consider the result of optimal interpolation (OI) projected onto the global EOF basis functions.
- 2: Split the case study region into overlapping  $P \times P$  patches, here,  $20 \times 20$  patches
- 3: For each patch position  $s$ , use the Analog Ensemble Kalman Smoother (AnEnKS) [13], 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. The assimilation is sequential and is performed each 3-days.
- 4: Reconstruct 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).
- 5: the reconstruction of fields  $X$  as  $\bar{X} + dX$ .

We may point out two important aspects in the implementation of the proposed patch-level AnDA setting:

- (step 3) 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.
- (step 4) 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 a patchwise EOF-based decomposition-reconstruction with a smaller patch-size (here,  $17 \times 17$  patches) to remove these blocky artifacts.



**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.

## 5. 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  $20 \times 20$  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  $17 \times 17$  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) is considered to be diagonal  $\mathbf{R} = \kappa^2 \mathbf{I}$  and  $\kappa = 0.001\text{m}$ .
- *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 100 km 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 [37]. Using an iterative reconstruction scheme, VE-DINEOF starts by filling the missing data with a first guess, here along-track pseudo-observation field for along-track data positions and  $\bar{X}$  for missing data positions. For each iteration, the resulting field is projected on the most significant EOF components calculated from the clean catalog data, then missing data positions are updated using pixels from the reconstructed new field. We run this iterative process until convergence. To make this algorithm comparable to the proposed PB-AnDA setting, we perform the reconstruction for each patch position then regroup the results as in PB-AnDA.
- *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 a diagonal observation covariance similar to as in PB-AnDA.

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 [24] (available at [github.com/rfablet/PB\\_ANDA](https://github.com/rfablet/PB_ANDA)). Optimal Interpolation was implemented on Matlab using [36]. Throughout the experiments, 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. These two metrics allow a good evaluation on image reconstruction capabilities and are widely used in missing data interpolation literature [38,39].

## 6. Results

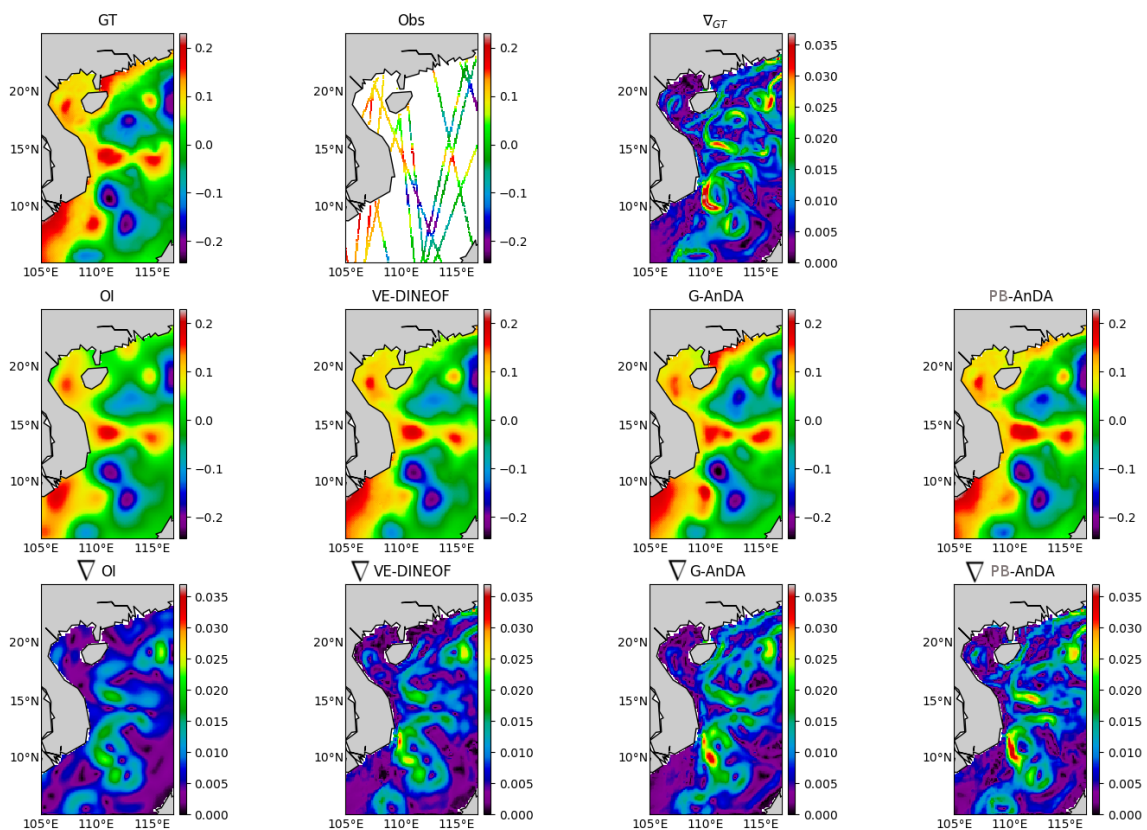
We evaluate the proposed PB-AnDA approach using the OSSE presented in Section 2. We perform a qualitative and quantitative comparison to state-of-the-art approaches. We first describe the parameter setting used for the PB-AnDA as well as benchmarked models, namely OI, an EOF-based approach [37] and a 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.

6.1. SLA Reconstruction from Noise-Free Along-Track Data

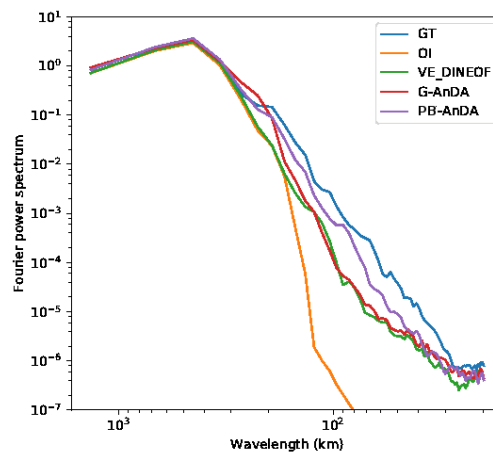
We first performed an idealized noise-free experiment, where the along-track observations were not contaminated with noise. The interpolation performances for this experiment are illustrated in Table 1. Our PB-AnDA algorithm significantly outperforms OI. More specifically, the locally-linear PB-AnDA 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) (meters), correlation statistics for Optimal Interpolation (OI), VE-DINEOF, G-AnDA and PB-AnDA with regard to the groundtruthed SLA fields. The relative gain with regard to OI is also shown in percentage. See Section 5 for the corresponding parameter settings.

Criterion	RMSE	Correlation	$\frac{RMSE_{OI} - RMSE}{RMSE_{OI}}$	
OI	0.026 ± 0.007	0.81 ± 0.08	-	
VE-DINEOF	0.023 ± 0.007	0.85 ± 0.07	11.53%	
G-AnDA	0.020 ± 0.006	0.89 ± 0.04	23.07%	
PB-AnDA	Locally-constant	0.014 ± 0.005	0.95 ± 0.03	46.15%
	Locally-Increment	0.014 ± 0.005	0.95 ± 0.03	46.15%
	Locally-Linear	<b>0.013 ± 0.005</b>	<b>0.96 ± 0.02</b>	<b>50.00%</b>



(a)  
Figure 4. Cont.



(b)

**Figure 4.** Reconstructed SLA fields (meters) using noise-free along-track observation using OI, VE-DINEOF, G-AnDA, PB-AnDA on the 54th day (24 February 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.

### 6.2. SLA Reconstruction from Noisy Along-Track Data

We also evaluated the proposed approach for noisy along-track data. Here, we ran two experiments with an additive zero-mean Gaussian noise applied to the simulated along-track data. We considered a diagonal noise covariance of  $\gamma^2 \mathbf{I}$  where  $\gamma = 0.01$  m (Experiment A) and  $\gamma = 0.03$  m (Experiment B) which was more close to the instrumental error of conventional altimeters. Given the resulting noisy along-track dataset, we applied the same methods as for the noise-free case study.

We ran PB-AnDA using different values for  $\kappa$ . For Experiment A, Table 2 shows that the minimum is reached using the true value of the error  $\kappa = 0.01$  m. While for Experiment B, Table 3 shows that the minimum is counter-intuitively reached again using value of the error  $\kappa = 0.01$  m with a negligible margin compared to the true value.

**Table 2.** Impact of standard variation of observation error  $\kappa$  in AnDA interpolation performance using noisy along-track data ( $\gamma = 0.01$  m): RMSE (meters) of AnDA interpolation for different values of parameter  $\kappa$ . For the same dataset, OI RMSE is 0.039.

$\kappa$	0.1	0.05	0.03	<b>0.01</b>	0.005	0.001	0.0001
$rmse_{PB-AnDA}$	0.035	0.030	0.028	<b>0.025</b>	0.025	0.029	0.044

**Table 3.** Impact of standard variation of observation error  $\kappa$  in AnDA interpolation performance using noisy along-track data ( $\gamma = 0.03$  m): RMSE (meters) of AnDA interpolation for different values of parameter  $\kappa$ . For the same dataset, OI RMSE is 0.066.

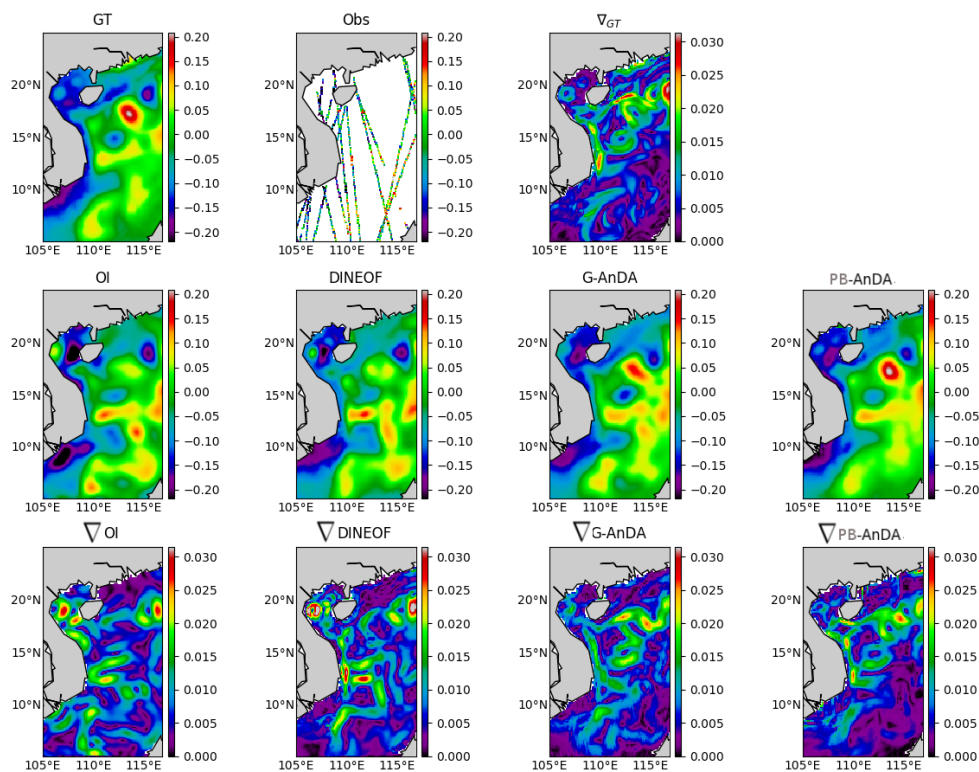
$\kappa$	0.1	0.05	0.03	<b>0.01</b>	0.005	0.001	0.0001
$rmse_{PB-AnDA}$	0.038	0.036	0.035	<b>0.0349</b>	0.037	0.046	0.076

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 Figure 5a compared with OI,

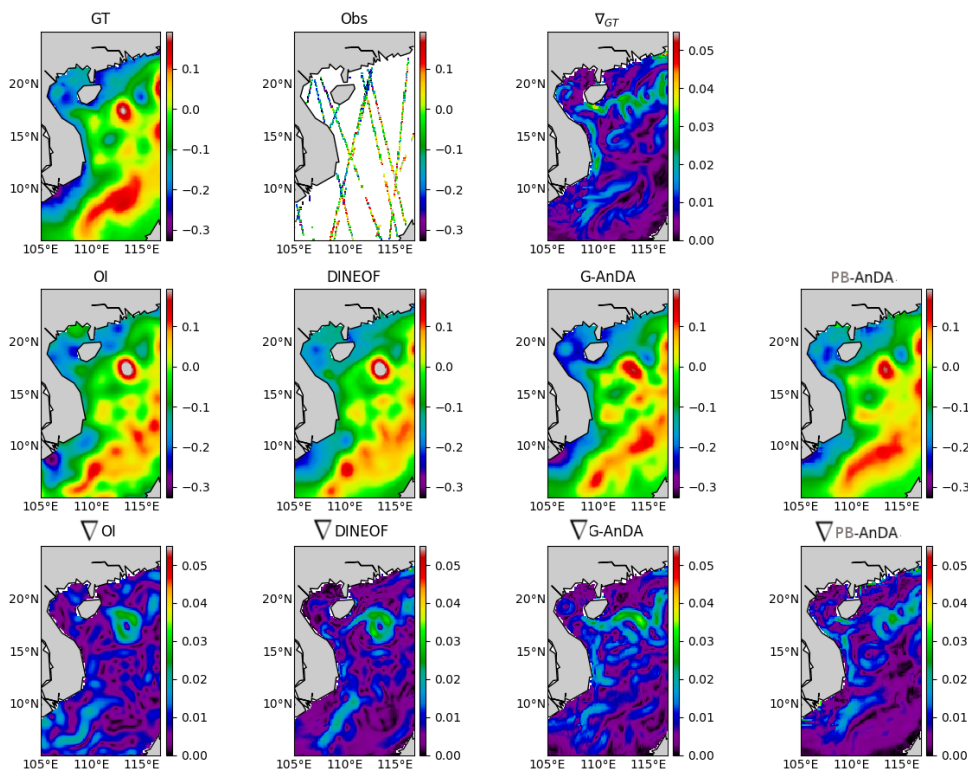
VE-DINEOF and G-AnDA. In Figure 5b, PB-AnDA also better reconstructs a larger-scale North-East structure, poorly sampled by along-track data and hence poorly interpolated by OI.

**Table 4.** SLA Interpolation performance for noisy along-track data: RMSE (meters) and correlation statistics for OI, VE-DINEOF, G-AnDA and PB-AnDA w.r.t. the groundtruthed SLA fields. The relative gain with regard to OI is also shown in percentage. See Section 5 for the corresponding parameter settings.

	Criterion	RMSE	Correlation	$\frac{RMSE_{OI} - RMSE}{RMSE_{OI}}$	
$\gamma = 0.01$ m	OI	$0.039 \pm 0.005$	$0.64 \pm 0.09$	-	
	VE-DINEOF	$0.035 \pm 0.005$	$0.68 \pm 0.09$	10.25%	
	G-AnDA	$0.030 \pm 0.005$	$0.78 \pm 0.06$	23.07%	
	PB-AnDA	Locally constant	$0.026 \pm 0.005$	$0.82 \pm 0.05$	33.33%
		Increment	$0.028 \pm 0.005$	$0.81 \pm 0.05$	28.20%
Local Linear		<b><math>0.0245 \pm 0.005</math></b>	<b><math>0.83 \pm 0.05</math></b>	<b>37.17%</b>	
$\gamma = 0.03$ m	OI	$0.066 \pm 0.006$	$0.41 \pm 0.09$	-	
	VE-DINEOF	$0.060 \pm 0.006$	$0.45 \pm 0.09$	9.09%	
	G-AnDA	$0.039 \pm 0.006$	$0.67 \pm 0.09$	40.90%	
	PB-AnDA	Locally constant	$0.035 \pm 0.006$	$0.688 \pm 0.064$	46.96%
		Increment	$0.036 \pm 0.006$	$0.656 \pm 0.07$	45.45%
Local Linear		<b><math>0.032 \pm 0.006</math></b>	<b><math>0.708 \pm 0.063</math></b>	<b>51.51%</b>	



(a)  
Figure 5. Cont.



(b)

**Figure 5.** Reconstruction of SLA fields (meters) from noisy along-track data using OI, VE-DINEOF, G-AnDA & PB-AnDA on day 225th (a) & 228th (b).

### 6.3. PB-AnDA Models with Auxiliary Variables

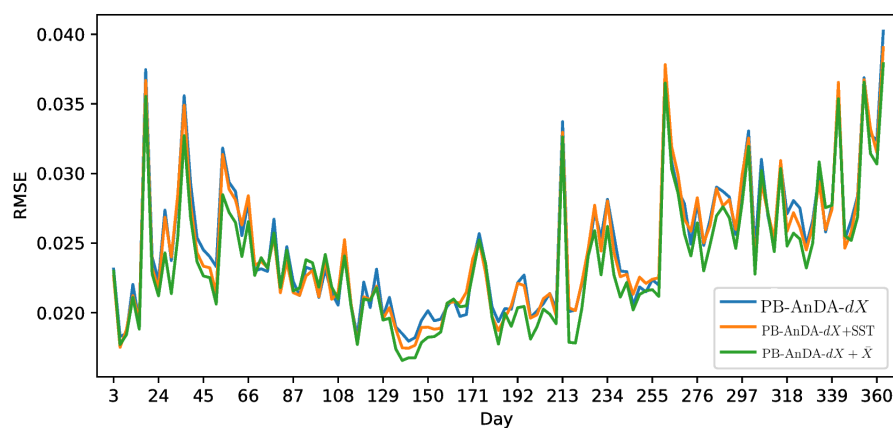
We further explore the flexibility of the analog setting to the use of additional geophysical variable information as explained in Section 4.2. Intuitively, we expect SLA fields to involve inter-scale dependencies as well as synergies with other tracers [19,40]. The use of auxiliary variables provide 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 groundtruthed 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 (meters) 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 [17], 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

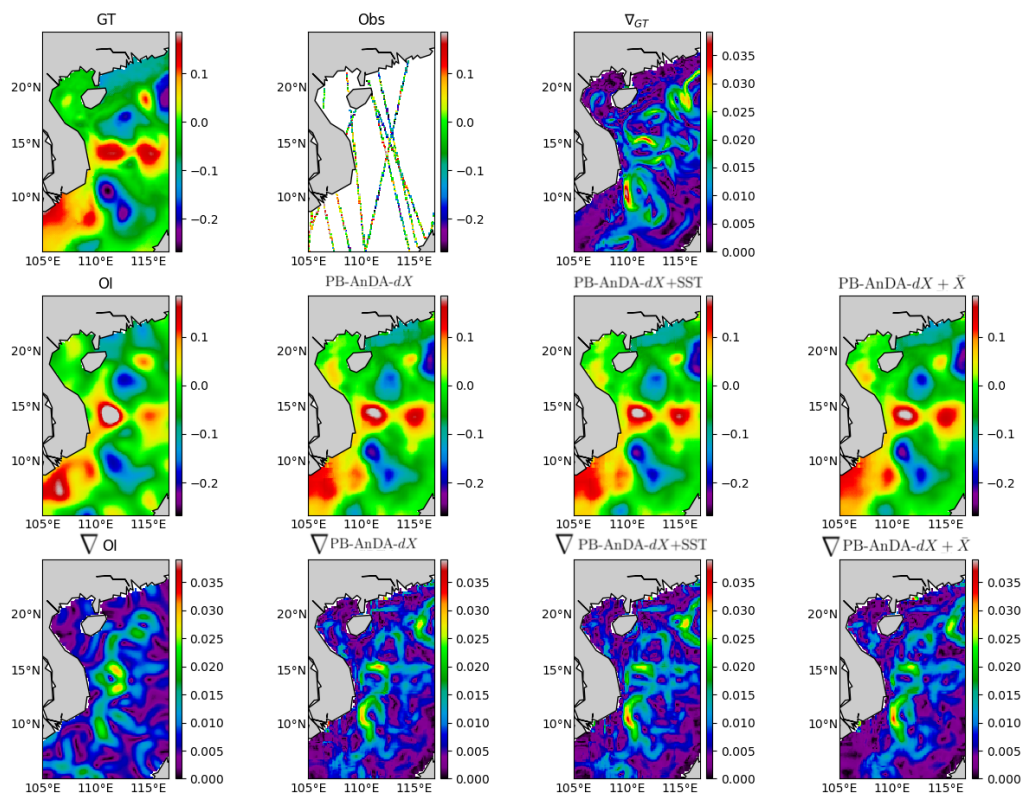
that PB-AnDA- $dX + \bar{X}$  always perform better than PB-AnDA- $dX$ . An example of the reconstruction is reported in Figure 7.

**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- $dX + SST$  (Sea Surface Temperature) 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  $\bar{X}^{GT}$ . The table shows the RMSE (meters) and correlation statistics.

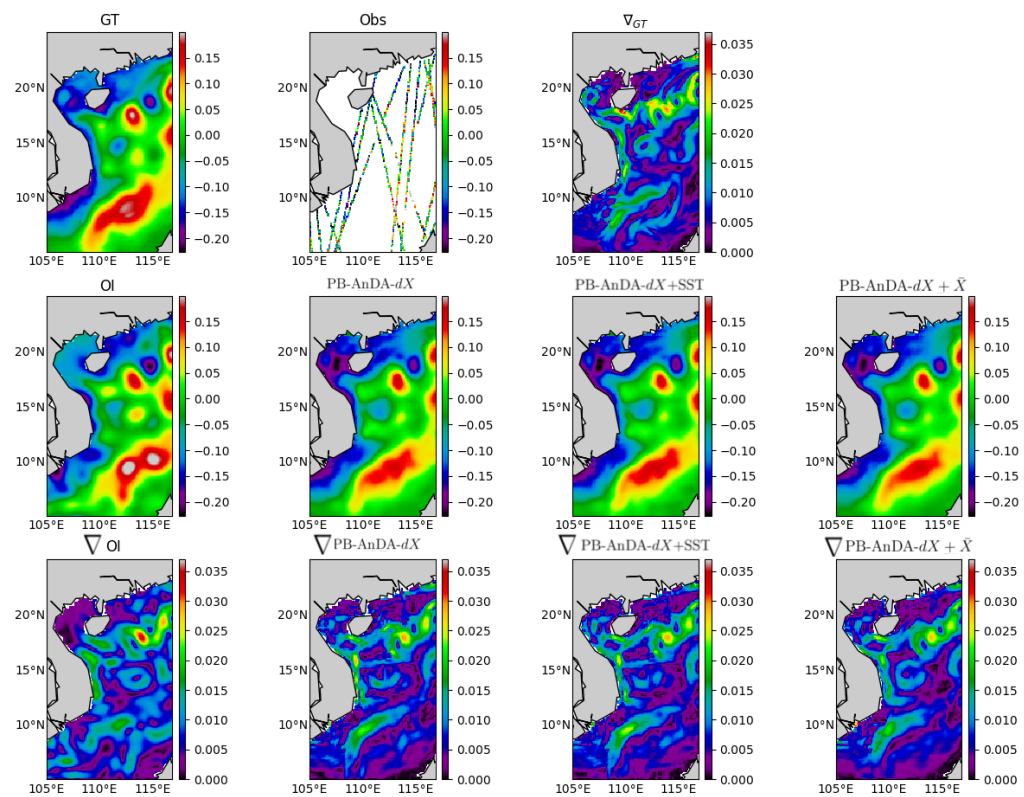
	PB-AnDA Model	RMSE	Correlation
$\gamma = 0.01$ m	PB-AnDA- $dX$	$0.025 \pm 0.005$	$0.83 \pm 0.05$
	PB-AnDA- $dX + SST$	$0.024 \pm 0.005$	$0.83 \pm 0.05$
	PB-AnDA- $dX + \bar{X}$	$0.023 \pm 0.005$	$0.84 \pm 0.05$
	PB-AnDA- $dX + \bar{X}^{GT}$	$0.021 \pm 0.004$	$0.87 \pm 0.04$
$\gamma = 0.03$ m	PB-AnDA-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



**Figure 6.** Daily RMSE (meters) 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 (meters) using PB-AnDA with different multivariate regression models on day 57th (a) & 237th (b).



## 7. Discussion

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 [18]. The combination of such physically-derived parameterizations to data-driven strategies appear to be a promising research direction. While we considered an OSSE to evaluate the proposed scheme, future work will investigate applications to real satellite-derived datasets, including the use of independent observation data such as surface drifters' track data to further assess the performance of the proposed algorithm.

The analog method is at the heart of this work as it is appealing by its implementation ease and its intuitive strategy, but it must not be seen as the only data-driven method adapted to the framework we presented. As long as a data-driven forecasting operator can be derived, other data-driven methods can be investigated [41,42]. One promising path is the use of neural networks, as they sparked off a series of breakthroughs in other fields [42–46]. While neural network based approaches can lead to better performances due to their superior regressing capabilities, two clear advantages of adopting analog methods are the fact that they do not need a time consuming training phase and that analog methods are easy to understand and interpret compared to black-box approaches such as neural networks.

The Analog Data Assimilation method as used in this work relies on several hyperparameters, assumptions and design choices. These considerations are discussed in the following:

- In this work, we used all the available data for the creation of the catalog, we expect that a rich dataset is important in increasing the likelihood of finding good analogs, yet a more thorough study is needed to assess the impact of reducing the temporal resolution of the dataset versus reducing the amount of the total available data. To compensate the computational impact of using a large dataset for the search for analogs, we used the FLANN (Fast Library for Approximate Nearest Neighbors) library as in [26]. This method makes use of tree-indexing techniques and is suitable for this kind of high dimensional applications [47].
- While we used a conditional Gaussian distribution in Equation (9), another alternative is the use of a conditional multinomial distribution. This resorts to sampling one of the analogs' successors as the forecast. Adopting this alternative would mean that we rely strongly on the archived catalog, so that forecasts of each ensemble member are actual elements of the catalog. This reduces the ability of the model to generate a rich variability of forecast scenes as done by the conditional Gaussian distribution.
- While we used a Gaussian kernel in Equation (10), other alternatives include cone kernels [10] which are more adapted to finding analogs in time series. The performance of our algorithm was slightly improved at the expense of more time execution and an additional hyperparameter to tune empirically, and we decided to prioritize simpler and efficient kernels. Regarding the scale parameter, the median was chosen due to its robustness to outliers.
- Although the AnEnKS was used in this work, a possible alternative is the use of an analog based Particle Smoother which drops Gaussian assumptions, however techniques based on particle filters need more ensemble members than their Ensemble Kalman Filter counterparts, thus causing a considerable increase in computational demands which was impractical for our application.

- We encourage the reader to refer to the discussion section in [24] for more insights about the rationale behind the use of patch-based representations and EOF-based dimensionality reduction.

Through the experiments conducted in this work, it was shown that the best performance was always reached using the locally-linear analog operator, which is in line with our previous findings [13,24]. An explanation for the superiority of this approach is that it better approximates locally the true dynamical model [48].

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 [49].

## 8. Materials and Methods

The experimental results presented in this work were obtained using the Python PB-AnDA toolbox made available by the authors at [https://github.com/rfablet/PB\\_ANDA](https://github.com/rfablet/PB_ANDA).

## 9. Conclusions

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 100 km) whereas the proposed patch-based analog setting successfully applies to the finer-scale range below 100 km. This is in agreement with the recent application of the analog data assimilation to the reconstruction of cloud-free SST fields [24]. 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 [17], we demonstrate that the additional use of large-scale SLA information may further improve the reconstruction performance for fine-scale structures.

We may state here the limitations of the present work and possible research avenues for the future. The experiments presented in this work were conducted on a numerical simulation derived dataset. A major future work direction would be then to apply the same procedure on real satellite-derived SLA contaminated with more complex noise models, then investigate the contribution of the use of numerical simulation datasets as catalogs. As combining multi-source datasets can also be challenging when using auxiliary variables relationships with SLA, an interesting experiment for example would be constructing a catalog with real SST observations combined with numerical simulation SLA datasets.

More efforts should be directed to assess the quality of the catalogs (spatio-temporal resolution, total years of measurements to consider, occurrence of rare events, etc.). Besides, building a good catalog can represent an opportunity for the use of neural networks based methods, and confronting these powerful regressors to our method is a promising future step. We also note that PB-AnDA can be a relevant candidate for the interpolation of other geophysical variables (e.g., Sea Surface Salinity, Chlorophyll concentrations, etc.) under the condition that they verify the set of assumptions made in this work.

**Author Contributions:** R.L., R.F. and G.C. conceived and designed the experiments; R.L. and M.S. created the datasets; R.L. and P.H.V. performed the experiments; R.L. and M.S. analyzed the data; T.F., B.C. contributed materials/analysis tools; R.L. and R.F. wrote the paper.

**Funding:** Redouane Lguensat is funded through a CNES (French Space Agency) postdoctoral grant. This work is also supported by ANR (Agence Nationale de la Recherche, grant ANR-13-MONU-0014), Labex Cominlabs (grant SEACS) and TeraLab (grant TIAMSEA).

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

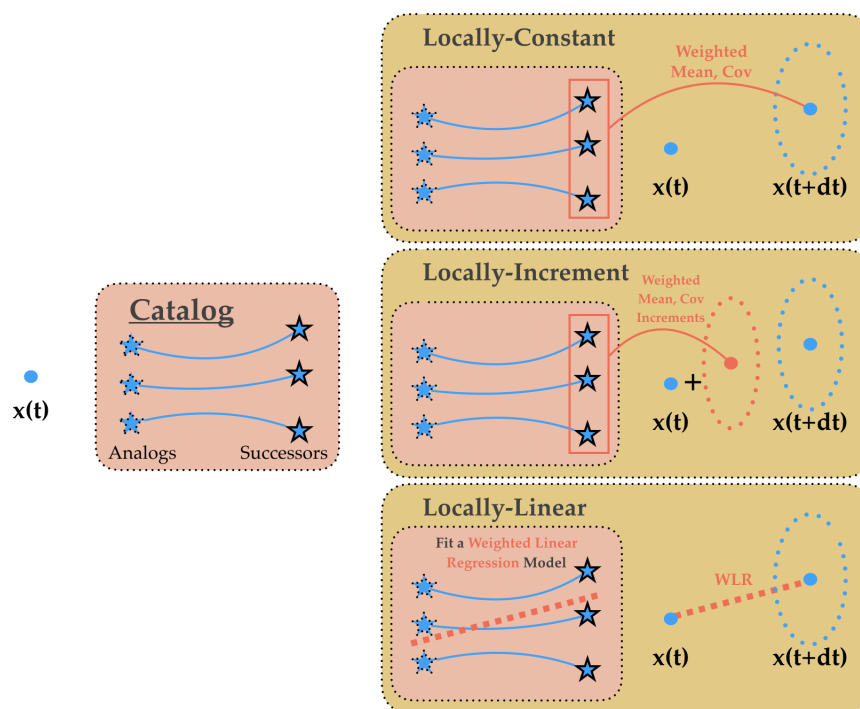
**Abbreviations**

The following abbreviations are used in this manuscript:

- AVISO+    Archiving, Validation et Interprétation des données des Satellites Océanographiques
- CMEMS    Copernicus Marine Environment Monitoring Service
- OSSE      Observation System Simulation Experiment
- OGCM      Ocean General Circulation Model
- OFES      OGCM for the Earth Simulation

**Appendix A. Analog Forecasting Operators**

In this appendix, we present the calculations needed for the three analog forecasting operators used in this work. An illustration is also given in Figure A1. Following [13], we give for each operator, the equations for  $\mu$  and  $\Sigma$  given  $\mathcal{A}$  the analogs of  $x(t - 1)$ , their successors  $\mathcal{S}$  and the corresponding weights  $\mathcal{K}_G$ :



**Figure A1.** Illustration of the three analog forecasting operator.

*Locally-Constant Operator*

$$\mu = \sum_{k=1}^K \mathcal{K}_G(x(t - 1), \mathcal{A}_k) \mathcal{S}_k(x(t - 1))$$

$$\Sigma = cov_{\mathcal{K}_G}(\mathcal{S}_k(x(t - 1)))_{k \in 1, K} \text{ where } cov_{\mathcal{K}_G} \text{ is a weighted covariance.}$$

### Locally-Incremental Operator

$$\begin{aligned}\tau_k(x(t-1)) &= \mathcal{S}_k(x(t-1)) - \mathcal{A}_k(x(t-1)) \\ \mu &= x(t-1) + \sum_{k=1}^K \mathcal{K}_G(x(t-1), \mathcal{A}_k) \tau_k(x(t-1)) \\ \Sigma &= \text{cov}_{\mathcal{K}_G}(x(t-1) + \tau_k(x(t-1)))_{k \in 1, K}\end{aligned}$$

### Locally-Linear Operator

Fitting a weighted least square between the  $K$  analogs and their successors we obtain slope  $\alpha(x(t-1))$  and intercept  $\beta(x(t-1))$  parameters, and residuals  $\xi_k(x(t-1))$  that lead us to  $\mu$  and  $\Sigma$ :

$$\begin{aligned}\xi_k(x(t-1)) &= \mathcal{S}_k(x(t-1)) - (\alpha(x(t-1))\mathcal{A}_k(x(t-1)) + \beta(x(t-1))) \\ \mu &= \alpha(x(t-1))x(t-1) + \beta(x(t-1)) \\ \Sigma &= \text{cov}((\xi_k(x(t-1)))_{k \in 1, K})\end{aligned}$$

## References

- Zhang, L.; Zhang, L.; Du, B. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 22–40. [[CrossRef](#)]
- Lary, D.J.; Alavi, A.H.; Gandomi, A.H.; Walker, A.L. Machine learning in geosciences and remote sensing. *Geosci. Front.* **2016**, *7*, 3–10. [[CrossRef](#)]
- Lorenz, E.N. Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.* **1969**, *26*, 636–646. [[CrossRef](#)]
- McDermott, P.L.; Wikle, C.K. A model-based approach for analog spatio-temporal dynamic forecasting. *Environmetrics* **2015**. [[CrossRef](#)]
- Comeau, D.; Giannakis, D.; Zhao, Z.; Majda, A.J. Predicting regional and pan-Arctic sea ice anomalies with kernel analog forecasting. *arXiv* **2017**, arXiv:1705.05228.
- Atencia, A.; Zawadzki, I. A Comparison of Two Techniques for Generating Nowcasting Ensembles. Part II: Analogs Selection and Comparison of Techniques. *Mon. Weather Rev.* **2015**, *143*, 2890–2908. [[CrossRef](#)]
- Delle Monache, L.; Eckel, F.A.; Rife, D.L.; Nagarajan, B.; Searight, K. Probabilistic weather prediction with an analog ensemble. *Mon. Weather Rev.* **2013**, *141*, 3498–3516. [[CrossRef](#)]
- Yiou, P. AnaWEGE: A weather generator based on analogues of atmospheric circulation. *Geosci. Model Dev.* **2014**, *7*, 531–543. [[CrossRef](#)]
- Hamill, T.M.; Whitaker, J.S. Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Weather Rev.* **2006**, *134*, 3209–3229. [[CrossRef](#)]
- Zhao, Z.; Giannakis, D. Analog Forecasting with Dynamics-Adapted Kernels. *arXiv* **2014**. arXiv: 1412.3831.
- Horton, P.; Jaboyedoff, M.; Obled, C. Global Optimization of an Analog Method by Means of Genetic Algorithms. *Mon. Weather Rev.* **2017**, *145*, 1275–1294. [[CrossRef](#)]
- Tandeo, P.; Ailliot, P.; Chapron, B.; Lguensat, R.; Fablet, R. The analog data assimilation: Application to 20 years of altimetric data. In Proceedings of the CI 2015: 5th International Workshop on Climate Informatics, Boulder, CO, USA, 24–25 September 2015; pp. 1–2.
- Lguensat, R.; Tandeo, P.; Ailliot, P.; Pulido, M.; Fablet, R. The Analog Data Assimilation. *Mon. Weather Rev.* **2017**, *145*, 4093–4107. [[CrossRef](#)]
- Hardman-Mountford, N.; Richardson, A.; Boyer, D.; Kreiner, A.; Boyer, H. Relating sardine recruitment in the Northern Benguela to satellite-derived sea surface height using a neural network pattern recognition approach. In *Progress in Oceanography: ENVIFISH: Investigating Environmental Causes of Pelagic Fisheries Variability in the SE Atlantic*; Elsevier Ltd.: Amsterdam, The Netherlands, 2003; Volume 59, pp. 241–255.
- Le Traon, P.; Nadal, F.; Ducet, N. An improved mapping method of multisatellite altimeter data. *J. Atmos. Ocean. Technol.* **1998**, *15*, 522–534. [[CrossRef](#)]
- Bretherton, F.P.; Davis, R.E.; Fandry, C. A technique for objective analysis and design of oceanographic experiments applied to MODE-73. In *Deep Sea Research and Oceanographic Abstracts*; Elsevier: Amsterdam, The Netherlands, 1976; Volume 23, pp. 559–582.
- Fablet, R.; Verron, J.; Mourre, B.; Chapron, B.; Pascual, A. Improving mesoscale altimetric data from a multi-tracer convolutional processing of standard satellite-derived products. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2518–2525. [[CrossRef](#)]

18. Ubelmann, C.; Klein, P.; Fu, L.L. Dynamic Interpolation of Sea Surface Height and Potential Applications for Future High-Resolution Altimetry Mapping. *J. Atmos. Ocean. Technol.* **2014**, *32*, 177–184. [[CrossRef](#)]
19. Klein, P.; Isern-Fontanet, J.; Lapeyre, G.; Rouillet, G.; Danioux, E.; Chapron, B.; Le Gentil, S.; Sasaki, H. Diagnosis of vertical velocities in the upper ocean from high resolution sea surface height. *Geophys. Res. Lett.* **2009**, *36*, L12603. [[CrossRef](#)]
20. Isern-Fontanet, J.; Chapron, B.; Lapeyre, G.; Klein, P. Potential use of microwave sea surface temperatures for the estimation of ocean currents. *Geophys. Res. Lett.* **2006**, *33*, L24608. [[CrossRef](#)]
21. Isern-Fontanet, J.; Shinde, M.; Andersson, C. On the Transfer Function between Surface Fields and the Geostrophic Stream Function in the Mediterranean Sea. *J. Phys. Oceanogr.* **2014**, *44*, 1406–1423. [[CrossRef](#)]
22. Turiel, A.; Sole, J.; Nieves, V.; Ballabrera-Poy, J.; Garcia-Ladona, E. Tracking oceanic currents by singularity analysis of Microwave Sea Surface Temperature images. *Remote Sens. Environ.* **2009**, in press. [[CrossRef](#)]
23. Turiel, A.; Nieves, V.; Garcia-Ladona, E.; Font, J.; Rio, M.H.; Larnicol, G. The multifractal structure of satellite sea surface temperature maps can be used to obtain global maps of streamlines. *Ocean Sci.* **2009**, *5*, 447–460. [[CrossRef](#)]
24. Fablet, R.; Viet, P.H.; Lguensat, R. Data-driven Models for the Spatio-Temporal Interpolation of satellite-derived SST Fields. *IEEE Trans. Comput. Imaging* **2017**. [[CrossRef](#)]
25. Buades, A.; Coll, B.; Morel, J.M. A non-local algorithm for image denoising. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR'05, San Diego, CA, USA, 20–25 June 2005; Volume 2, pp. 60–65. [[CrossRef](#)]
26. Fablet, R.; Huynh Viet, P.; Lguensat, R.; Horrein, P.H.; Chapron, B. Spatio-Temporal Interpolation of Cloudy SST Fields Using Conditional Analog Data Assimilation. *Remote Sens.* **2018**, *10*, 310. [[CrossRef](#)]
27. Masumoto, Y.; Sasaki, H.; Kagimoto, T.; Komori, N.; Ishida, A.; Sasai, Y.; Miyama, T.; Motoi, T.; Mitsudera, H.; Takahashi, K.; et al. A fifty-year eddy-resolving simulation of the world ocean: Preliminary outcomes of OFES (OGCM for the Earth Simulator). *J. Earth Simul.* **2004**, *1*, 35–56.
28. Sasaki, H.; Nonaka, M.; Masumoto, Y.; Sasai, Y.; Uehara, H.; Sakuma, H. An eddy-resolving hindcast simulation of the quasi-global ocean from 1950 to 2003 on the Earth Simulator. In *High Resolution Numerical Modelling of the Atmosphere and Ocean*; Springer: Berlin/Heidelberg, Germany, 2008.
29. Shaw, P.T.; Chao, S.Y.; Fu, L.L. Sea surface height variations in the South China Sea from satellite altimetry. *Oceanol. Acta* **1999**, *22*, 1–17. [[CrossRef](#)]
30. Bocquet, M.; Pires, C.A.; Wu, L. Beyond Gaussian statistical modeling in geophysical data assimilation. *Mon. Weather Rev.* **2010**, *138*, 2997–3023. [[CrossRef](#)]
31. Ide, K.; Courtier, P.; Ghil, M.; Lorenc, A. Unified notation for data assimilation: operational, sequential and variational. *Practice* **1997**, *75*, 181–189.
32. Asch, M.; Bocquet, M.; Nodet, M. *Data Assimilation: Methods, Algorithms, and Applications*; Fundamentals of Algorithms; SIAM: Philadelphia, PA, USA, 2016.
33. De Mey, P.; Robinson, A.R. Assimilation of altimeter eddy fields in a limited-area quasi-geostrophic model. *J. Phys. Oceanogr.* **1987**, *17*, 2280–2293. [[CrossRef](#)]
34. Gandin, L. Objective analysis of meteorological fields. By L. S. Gandin. Translated from the Russian. Jerusalem (Israel Program for Scientific Translations), 1965. Pp. vi, 242: 53 Figures; 28 Tables. £4 1s. 0d. *Q. J. R. Meteorol. Soc.* **1966**, *92*, 447. [[CrossRef](#)]
35. Lorenc, A.C. Analysis methods for numerical weather prediction. *Q. J. R. Meteorol. Soc.* **1986**, *112*, 1177–1194. [[CrossRef](#)]
36. Escudier, R.; Bouffard, J.; Pascual, A.; Poulain, P.M.; Pujol, M.I. Improvement of coastal and mesoscale observation from space: Application to the northwestern Mediterranean Sea. *Geophys. Res. Lett.* **2013**, *40*, 2148–2153. [[CrossRef](#)]
37. Ping, B.; Su, F.; Meng, Y. An Improved DINEOF Algorithm for Filling Missing Values in Spatio-Temporal Sea Surface Temperature Data. *PLoS ONE* **2016**, *11*, e0155928. [[CrossRef](#)]
38. Gerber, F.; de Jong, R.; Schaepman, M.E.; Schaepman-Strub, G.; Furrer, R. Predicting Missing Values in Spatio-Temporal Remote Sensing Data. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2841–2853. [[CrossRef](#)]
39. Fablet, R.; Rousseau, F. Missing data super-resolution using non-local and statistical priors. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 27–30 September 2015; pp. 676–680. [[CrossRef](#)]

40. Bernard, D.; Boffetta, G.; Celani, A.; Falkovich, G. Inverse Turbulent Cascades and Conformally Invariant Curves. *Phys. Rev. Lett.* **2007**, *98*, 024501. [[CrossRef](#)]
41. Moazen-zadeh, R.; Mohammadi, B.; Shamshirband, S.; Chau, K.W. Coupling a firefly algorithm with support vector regression to predict evaporation in northern Iran. *Eng. Appl. Comput. Fluid Mech.* **2018**, *12*, 584–597. [[CrossRef](#)]
42. Faizollahzadeh Ardabili, S.; Najafi, B.; Shamshirband, S.; Minaei Bidgoli, B.; Deo, R.C.; Chau, K.W. Computational intelligence approach for modeling hydrogen production: A review. *Eng. Appl. Comput. Fluid Mech.* **2018**, *12*, 438–458. [[CrossRef](#)]
43. Yaseen, Z.M.; Sulaiman, S.O.; Deo, R.C.; Chau, K.W. An enhanced extreme learning machine model for river flow forecasting: state-of-the-art, practical applications in water resource engineering area and future research direction. *J. Hydrol.* **2018**. [[CrossRef](#)]
44. Taormina, R.; Chau, K.W. Data-driven input variable selection for rainfall-runoff modeling using binary-coded particle swarm optimization and Extreme Learning Machines. *J. Hydrol.* **2015**, *529*, 1617–1632. [[CrossRef](#)]
45. Taherei Ghazvinei, P.; Hassanpour Darvishi, H.; Mosavi, A.; Yusof, K.b.W.; Alizamir, M.; Shamshirband, S.; Chau, K.w. Sugarcane growth prediction based on meteorological parameters using extreme learning machine and artificial neural network. *Eng. Appl. Comput. Fluid Mech.* **2018**, *12*, 738–749. [[CrossRef](#)]
46. Wu, C.; Chau, K. Rainfall runoff modeling using artificial neural network coupled with singular spectrum analysis. *J. Hydrol.* **2011**, *399*, 394–409. [[CrossRef](#)]
47. Muja, M.; Lowe, D.G. Scalable nearest neighbor algorithms for high dimensional data. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 2227–2240. [[CrossRef](#)]
48. Cleveland, W.S. Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.* **1979**, *74*, 829–836. [[CrossRef](#)]
49. Gaultier, L.; Ubelmann, C.; Fu, L.L. The Challenge of Using Future SWOT Data for Oceanic Field Reconstruction. *J. Atmos. Ocean. Technol.* **2015**, *33*, 119–126. [[CrossRef](#)]



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).