SPATIO-TEMPORAL INTERPOLATION OF ALTIMETER-DERIVED SSH FIELDS USING ANALOG DATA ASSIMILATION: A CASE-STUDY IN THE SOUTH CHINA SEA

Redouane Lguensat[*1]*, Miao Sun*[2]*, Ge Chen*[2,3]*, Fenglin Tian*[2]*, Ronan Fablet*[1]

[1] IMT Atlantique, LabSTICC, Université Bretagne Loire; Brest, France
[2] College of Information Science and Engineering, Ocean University of China, Qingdao, PR China
[3] Qingdao National Laboratory for Marine Science and Technology, PR China
[*] Contact: redouane.lguensat@imt-atlantique.fr

ABSTRACT

The reconstruction of high-resolution gridded altimetry maps from irregularly sampled along-track data remains a key challenge in ocean remote sensing science. Operational products use optimal Interpolation (OI) techniques, which may not deal with nonlinear dynamics at short space-time scales. Here, we investigate an analog data assimilation scheme to improve the reconstruction of fine-scale structures. The analog data assimilation combines an ensemble Kalman model and a dataset of exemplars issued from high-resolution numerical simulations to perform an exemplar-based spatio-temporal interpolation of along-track data. As a case-study, we consider a region in the South China Sea and demonstrate the proposition analog data assimilation outperforms the classical OI by about $\simeq 20\%$ in terms of mean square reconstruction error.

Index Terms— Satellite Altimetry, Optimal Interpolation, Analog Data Assimilation.

1. INTRODUCTION

Optimal Interpolation (or objective analysis) aims at determining the best estimation of a field using irregularly spaced observations. Under the assumption that the field is Gaussian and its covariance structure is known, the technique uses the Gauss-Markov theorem to minimize the mean square error of the estimate [1]. OI was introduced to ocean science in [2], and is the operational tool [3, 4] for ocean altimetry mapping. Still, OI suffers from two major limitations:

- The Gauss-Markov estimate optimality depends on the correctness of the Gaussianity assumption and of the time and length scales used for the space-time correlation function.
- While the Gaussianity assumption is mostly relevant for large-scale structure, this assumption does not hold for small-scale structures, which makes OI more relevant for the large-scale (low-frequency) components of the SSH fields.

The main idea of this work consists in expressing a Sea Level Anomaly (SLA) map as a sum of two scale components, namely, large and fine scales. The large scale component is reconstructed from along-track data using a classical OI, whereas we investigate a non-parametric exemplar-based method for the fine-scale component. The past twenty years have witnessed an important growth of ocean satellite data originating from satellite networks, model simulations, *in situ* data, etc. This amount of available data supports the development of machine learning and data-driven methods for ocean remote sensing applications [5, 6, 7].

Here, we investigate such a data-driven approach, namely analog data assimilation (AnDA) framework introduced recently [8]. AnDA combines analog forecasting methods [9] and stochastic data assimilation algorithms. In [8], we considered an application to Lorenz 63 and Lorenz 96 models, which involve rather low-dimensional state space ($<$100). By contrast, we consider here a more challenging problem as gridded SSH field involves significantly higher-dimensional state space, typically above 10^5. Our key contribution lies both in considering a more complex application and in dealing with the SLA mapping problem using AnDA within a multiscale approach. Hereinafter, the method we introduce is called Multiscale Analog Data Assimilation (MS-AnDA).

This paper is organized as follows. Section 2 describes the considered data. Section 3 presents the multiscale analog data assimilation method. Numerical experiments are reported in Section 4 and we finally comment our findings and state our future work in Section 5.

2. DATA PREPARATION

For evaluation purposes, we consider an Observation System Simulation Experiment (OSSE). Using numerical simulations, we create along-track data using the real along-track sampling patterns in 2014 with four altimeters. We describe below the data preparation setup.

2.1. Model simulation data

We consider a 50 years 3-daily SSH time series from 1962 to 2012 from the Ocean General Circulation Model (OGCM) for the Earth Simulator (OFES [10, 11]). The coverage of the model is 75°S-75°N with a horizontal resolution of $1/10°$. Our region of interest is in the South China Sea (105°E to 117°E, 5°N to 25°N).

We use these numerical simulations to run an OSSE, which relies on the sampling of pseudo along-track data as described in the next section.

2.2. Along track data

Along track data represent the direct measurements of SSH from each satellite on its orbit around the globe. In this work, we use 2014 along-track data positions from 4 satellites (Jason2, Cryosat2, Saral/AltiKa, HY-2A) distributed by Copernicus Marine and Environment Monitoring Service (CMEMS) `http://marine.copernicus.eu`. We use the resulting along-track position dataset to generate pseudo along-track data from OGCM numerical simulations.

3. MULTISCALE ANALOG DATA ASSIMILATION

Let us denote by X the gridded (interpolated) SSH field. As stated in the introduction we represent the SSH field using a two-scale representation. Formally, we express our model as follows:

$$X = \bar{X} + dX_1 + \xi \tag{1}$$

where $\bar{X}$ refers to the large-scale component, dX_1 to the fine-scale component and ξ to the unresolved scales.

From along-track data, we first reconstruct $\bar{X}$ using the classical optimal interpolation algorithm. The MS-AnDA is then applied to the detail field dX_1 as explained in the next paragraph.

The MS-AnDA algorithm is an extension of the analog data assimilation (AnDA) algorithm. AnDA is a data-driven data assimilation scheme *i.e.* the dynamical model is learned from data, contrarily to classical data assimilation where a physical model of the dynamics in needed. The particularity of the MS-AnDA lies in the combination of a patch-based and EOF-based representation (EOF: Empirical Orthogonal functions, also known as Principal Component Analysis), which makes the algorithm more suited to high dimensional problems as in this study.

Given the OI-interpolated large-scale component $\bar{X}$ of Eq.1, we consider a patch-based representation of residual field $X - \bar{X}$, referred to as dX. Let us denote by $\mathcal{P}_s$ the $P \times P$ patch centered at site i. The reconstruction of field $dX = X - \bar{X}$ resorts to the following model:

$$\left\{ \begin{array}{rcl} dX(\mathcal{P}_i, t) & = & \mathcal{A}(dX(\mathcal{P}_i, t-1)) + \eta(\mathcal{P}_i, t), \forall i \\ Y(\mathcal{P}_i, t) & = & \mathcal{H}(X(\mathcal{P}_i, t)) + \epsilon(\mathcal{P}_i, t) \end{array} \right. \tag{2}$$

where $Z(\mathcal{P}_i, t)$ refers to the $P \times P$ patch centered in i for field Z at time t. Following AnDA scheme [8], $\mathcal{A}$ is an analog dynamical model [9]. It retrieves exemplars in a reference catalog similar to the current state at time $t-1$ to sample forecasts at time t. Y is the observation field, *i.e.* along-track data. $\mathcal{H}$ is the observation operator which accounts for the irregular sampling of the along-track data. η and ϵ are independent Gaussian centered noises that represents uncertainty in the model and observation equations.

The key component of the AnDA is the selected analog forecasting strategy. We use here an incremental forecasting strategy as follows. Let us suppose we want to forecast state $dX(\mathcal{P}_i, t-1)$, to simplify reading, let denote it by dX^i_{t-1}. We first find the K nearest neighbors (or analogs) of dX^i_{t-1} in our database. We note these analogs by $a_k(dX^i_{t-1})_{k\in[\![1..K]\!]}$, and their coressponding forecasts (successors) by $s_k(dX^i_{t-1})_{k\in[\![1..K]\!]}$. Forecasting dX^i_{t-1} resorts to adding to dX^i_{t-1} a weighted mean of the K increments τ_k *i.e.* differences between analogs and successors $\tau_k(dX^i_{t-1}) = s_k(dX^i_{t-1}) - a_k(dX^i_{t-1})$. If we note by ω the weights, $\mathcal{A}$ and η from the model equation in Eq.2 are then expressed as follows:

$$\begin{array}{rcl} \mathcal{A}(dX^i_{t-1}) & = & dX^i_{t-1} + \sum_{k=1}^K \omega_k(dX^i_{t-1})\tau_k(dX^i_{t-1}) \\ & = & \sum_{k=1}^K \omega_k(dX^i_{t-1})(dX^i_{t-1} + \tau_k(dX^i_{t-1})) \end{array}$$

$\eta(t)$ is a Gaussian centered noise of covariance matrix $\Sigma_{LI} = cov_\omega((dX^i_{t-1} + \tau_k(dX^i_{t-1}))_{k\in[\![1,K]\!]})$, where cov_ω stands for weighted covariance. The weight ω_k associated to the successor s_k depends on the distance between the state dX^i_{t-1} and its analog a_k. In this work, a Gaussian kernel is used as follows:

$$\omega_k \propto \exp(-\frac{\|dX^i_{t-1} - a_k\|^2}{\sigma^2}) \tag{3}$$

Given that state, the analogs and successors are all patch images of size $P \times P$, the search of nearest neighbors is more likely to fail, and the computational complexity of the problem overall is high. Here comes the EOF decomposition, to reduce the dimensionality of the problem from a $P \times P$ dimension to few N_{EOF} dimensions, here $N_{EOF} = 10$. The mathematical resolution of Eq.(2) for every region patch is then performed using a data-driven ensemble-based data assimilation algorithm, namely the Analog Ensemble Kalman Smoother (AnEnKS). The AnEnKS represents a version of the classical Ensemble Kalman Filter and Smoother (EnKF/EnKS) adapted to analog based data assimilation. The reader is invited to see [8] for more details. As stated early in the paragraph, we consider overlapping regions patches, however, to reduce the computational complexity we do not consider all possible positions, a stride of $str = 15$ is used to move from one patch to the other either horizontally or vertically.

4. NUMERICAL EXPERIMENTS

All implementations were run under Matlab. We use optimal interpolation code from [12], and the Analog Data Assimilation toolbox [8] (Python Library can be found at `https://github.com/ptandeo/AnDA` and the Matlab toolbox can be obtained by sending an email to the corresponding author).

Model simulation data is split as follows. The first 49 years of the time series are used for training, that is to say to build a catalog of analogs and successors. We apply the proposed approach to 2012 data. The true SSH field from the numerical simulation provides a ground truth to evaluate the relevance of the spatio-temporal interpolation in terms of root mean square error. In our experiments, we compare three schemes: (i) a OI interpolation using a Gaussian covariance model with 10-day and 100km correlation lengths, (ii) the direct application of the AnDA using a 10-component EOF decomposition of the whole SSH field *i.e.* no patch-based representation (OI+MS-AnDA$_{noP}$), (iii) the proposed patch-based MS-AnDA. For both AnDA schemes, we use 50 analogs in the analog forecasting steps. The patch-based representation exploits 20×20 patches and retains 10 EOFs. Table 1 reports the RMSE values obtained by the compared algorithms. Whereas we only report a marginal improvement of OI+MS-AnDA$_{noP}$ compared to OI (RMSE values of 0.033 vs. 0.035), the proposed AnDA models leads to a gain of about 20% (RMSE values of 0.028 vs. 0.035).

	OI	OI+MS-AnDA$_{noP}$	OI+MS-AnDA
RMSE	0.035	0.033	**0.028**

Table 1: RMSE comparison (in meters)

Figure 1 illustrates two examples of the results obtained using the proposed method. The error maps (second row) show a reduction of the error when adding the MS-AnDA reconstructed fine scale component. The performance of both algorithms is highlighted in the gradient magnitude maps (third row) where we see clearly on one hand the smoothing effect in OI result, and in the other hand, the gradient enhancement and the structures that made appearance in case of OI+MS-AnDA.

5. CONCLUSION AND FUTURE WORK

This study was motivated by the following question: How can data-driven methods (and especially analog methods) help improving SLA mapping using optimal interpolation? We address this question by proposing a multiscale based method where the fine scale component of the SLA maps is reconstructed using a data-driven data assimilation method. Given a database of model simulations we use a K-NN based method coming with ensemble-based filtering in order to re-

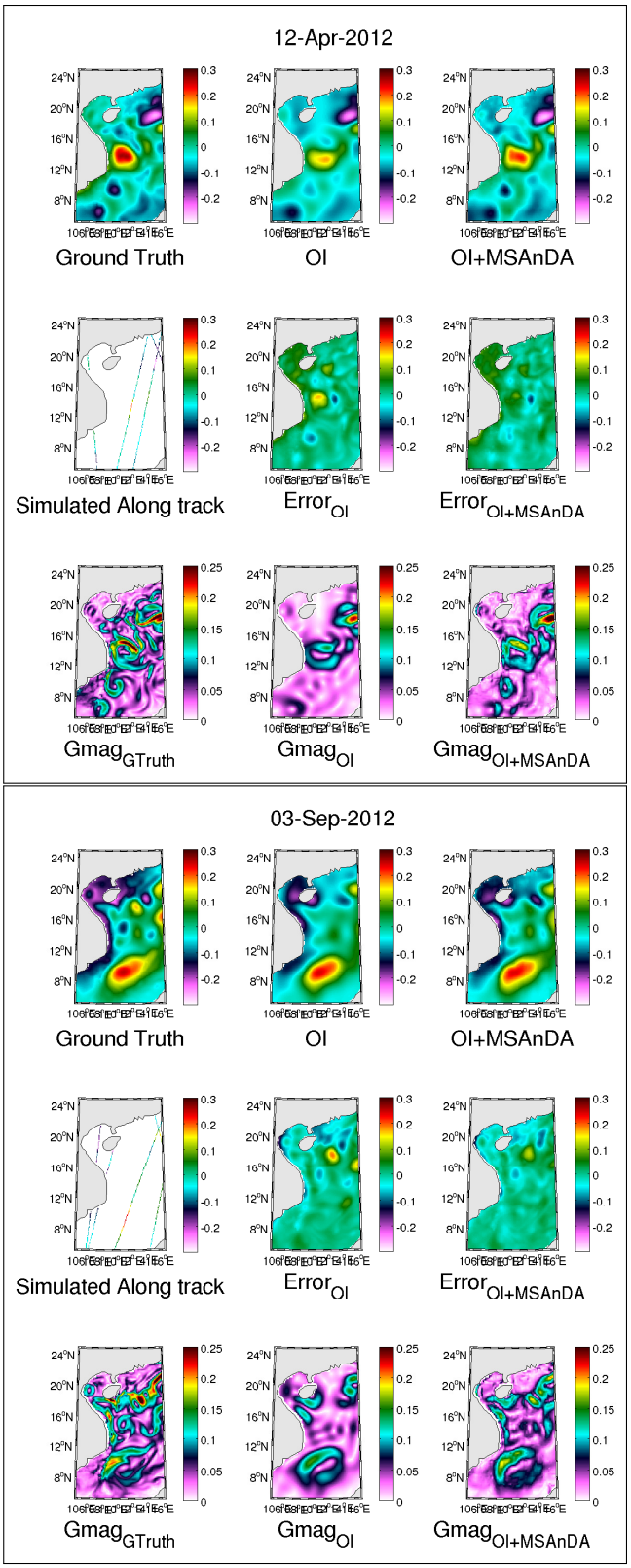

Fig. 1: The result of the proposed method for 2 different days. From left to right: (first row) ground truth, OI and OI+MS-AnDA, (second row) simulated along track, error maps using the difference between OI then OI+MS-AnDA and ground truth, (third row) gradient magnitude maps of ground truth, OI and OI+MS-AnDA.

trieve the lost information. The proposed method improves OI reconstruction in a reasonable time providing that we use patch based representation and EOF based dimensionality reduction.

While the reconstruction of dX_1 is done independently of $\bar{X}$ in this work, we may investigate an additional conditioning by $\bar{X}$. Future work may also explore other analog forecasting strategies: i) investigating better choices for the kernel by selecting more dynamically-adapted ones as used in [13], ii) making profit of the recent advancement in artificial neural networks (ANN) for forecasting and designing algorithms that can integrate ANN based forecasting techniques in the MS-AnDA paradigm. One could also explore the synergy of SSH fields and other satellite-derived sea surface fields, such as SST and ocean colour, to further improve the reconstruction of SSH fields. We think that this work and the directions of research we suggest will help in bringing more interest in applied machine learning to ocean altimetry mapping and offers some perspectives that are worthy of investigation.

6. 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).

7. REFERENCES

[1] Lee-Lueng Fu and Anny Cazenave, *Satellite altimetry and earth sciences: a handbook of techniques and applications*, vol. 69, Academic Press, 2000.

[2] Francis P Bretherton, Russ E Davis, and CB Fandry, "A technique for objective analysis and design of oceanographic experiments applied to mode-73," in *Deep Sea Research and Oceanographic Abstracts*. Elsevier, 1976, vol. 23, pp. 559–582.

[3] PY Le Traon, F Nadal, and N Ducet, "An improved mapping method of multisatellite altimeter data," *Journal of Atmospheric and Oceanic Technology*, vol. 15, no. 2, pp. 522–534, 1998.

[4] AVISO, "SSALTO/DUACS user handbook:(M) SLA and (M) ADT near-real time and delayed time products," 2009.

[5] David J Lary, Amir H Alavi, Amir H Gandomi, and Annette L Walker, "Machine learning in geosciences and remote sensing," *Geoscience Frontiers*, vol. 7, no. 1, pp. 3–10, 2016.

[6] Liangpei Zhang, Lefei Zhang, and Bo Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 22–40, 2016.

[7] Redouane Lguensat, Pierre Tandeo, Ronan Fablet, and René Garello, "Spatio-temporal interpolation of sea surface temperature using high resolution remote sensing data," in *Oceans-St. John's, 2014*. IEEE, 2014, pp. 1–4.

[8] Redouane Lguensat, Pierre Tandeo, Pierre Ailliot, Manuel Pulido, and Ronan Fablet, "The Analog Data Assimilation," 2016 (submitted).

[9] Edward N Lorenz, "Deterministic nonperiodic flow," *Journal of the atmospheric sciences*, vol. 20, no. 2, pp. 130–141, 1963.

[10] Yukio Masumoto, Hideharu Sasaki, Takashi Kagimoto, Nobumasa Komori, Akio Ishida, Yoshikazu Sasai, Toru Miyama, Tatsuo Motoi, Humio Mitsudera, Keiko Takahashi, et al., "A fifty-year eddy-resolving simulation of the world ocean: Preliminary outcomes of ofes (ogcm for the earth simulator)," *J. Earth Simulator*, vol. 1, pp. 35–56, 2004.

[11] Hideharu Sasaki, Masami Nonaka, Yukio Masumoto, Yoshikazu Sasai, Hitoshi Uehara, and Hirofumi Sakuma, *An Eddy-Resolving Hindcast Simulation of the Quasiglobal Ocean from 1950 to 2003 on the Earth Simulator*, pp. 157–185, Springer New York, New York, NY, 2008.

[12] Romain Escudier, Jérôme Bouffard, Ananda Pascual, Pierre-Marie Poulain, and Marie-Isabelle Pujol, "Improvement of coastal and mesoscale observation from space: Application to the northwestern mediterranean sea," *Geophysical Research Letters*, vol. 40, no. 10, pp. 2148–2153, 2013.

[13] Zhizhen Zhao and Dimitrios Giannakis, "Analog forecasting with dynamics-adapted kernels," *Nonlinearity*, vol. 29, no. 9, pp. 2888, 2016.