remote sensing

MDPI

Article

Spatio-Temporal Interpolation of Cloudy SST Fields Using Conditional Analog Data Assimilation

Ronan Fablet [1,*] Phi Huynh Viet [1], Redouane Lguensat [1], Pierre-Henri Horrein [1] and Bertrand Chapron [2]

[1] IMT Atlantique, Lab-STICC, UBL, Brest, 29238, France; vietphi3892@gmail.com (P.H.V.); redouane.lguensat@imt-atlantique.fr (R.L.); ph.horrein@imt-atlantique.fr (P.-H.H.)

[2] Ifremer, LOPS, Brest, 29200, France; bchapron@ifremer.fr

* Correspondence: ronan.fablet@imt-atlantique.fr; Tel.: +33-229-001-287

Received: 27 November 2017; Accepted: 6 February 2018; Published: 17 February 2018

Abstract: The ever increasing geophysical data streams pouring from earth observation satellite missions and numerical simulations along with the development of dedicated big data infrastructure advocate for truly exploiting the potential of these datasets, through novel data-driven strategies, to deliver enhanced satellite-derived gapfilled geophysical products from partial satellite observations. We here demonstrate the relevance of the analog data assimilation (AnDA) for an application to the reconstruction of cloud-free level-4 gridded Sea Surface Temperature (SST). We propose novel AnDA models which exploit auxiliary variables such as sea surface currents and significantly reduce the computational complexity of AnDA. Numerical experiments benchmark the proposed models with respect to state-of-the-art interpolation techniques such as optimal interpolation and EOF-based schemes. We report relative improvement up to 40%/50% in terms of RMSE and also show a good parallelization performance, which supports the feasibility of an upscaling on a global scale.

Keywords: ocean remote sensing data; data assimilation; optimal interpolation; analog models; multi-scale decomposition; patch-based representation

1. Introduction

Long records of high-resolution Sea Surface Temperature (SST) are of high importance for a wide range of applications including among others weather and climate forecasting, ocean-atmosphere exchanges, the monitoring of tropical cyclones [1]. SST is an example of essential variables derived from remote sensing data [2–6], which play a critical role in climate models as well as numerical weather forecasts. SST field time series are for instance among the key satellite-derived data assimilated in ocean-atmosphere models [7,8] and hurricane dynamics [9]. Spaceborne sensors provide invaluable data to reconstruct satellite-derived high-resolution SST fields (typically, up to a few kilometers) on a global scale. Such SST fields may however comprise high rates of missing data. Optical sensors [10] may depict the highest missing data rates, as they cannot sense the ocean surface through clouds. Though less sensitive to atmospheric conditions [11], radiometers are also affected by thick clouds and heavy rain conditions.

The reconstruction of gap-free high-resolution SST fields from satellite-derived SST measurement has long been a critical issue [12–18]. Operational products typically rely on the Optimal Interpolation (OI). Amon others, cloud-free OSTIA [12], ODYSSEA [19], AMSR-E [17] products are examples of operational products which rely on OI. It produces the Best Linear Unbiased Estimator (BLUE) of the field given irregularly sampled observations. This model-driven approach requires selecting a covariance prior of the SST fields, most often exponential and Gaussian covariance models [12,18]. The parameterization of this covariance prior involves a trade-off between the size of the gaps to be

filled and the fine-scale variability of the SST fields. Physically-driven data assimilation models [20] may outperform OI if relevant dynamical priors can be defined [21]. The trade-off to be considered between the complexity and genericity of this physical prior remains however complex, especially when considering the assimilation of a single sea surface tracer as SST.

Besides model-driven schemes, the ever increasing availability of satellite-derived data and of simulation data from high-resolution ocean models has paved the way for the development of data-driven methods. EOF-based models were among the early and perhaps most popular data-driven methods applied to the reconstruction of SST fields from cloudy SST data [14,15,22] as well of other sea surface tracers such as ocean colour [22]. EOF-based approaches are particularly appealing for ocean remote sensing as they relate to a model of the covariance structure of the considered fields and may adapt to any type of geometry of missing data and interpolation grid. Their use is also motivated by their ability to decompose the spatiotemporal variability of the sea surface fields according to different modes, which may be interpreted geophysically. A renewed interest can also be noticed for analog schemes and applications to forecasting and assimilation issues [23,24]. Analog schemes, proposed a long time ago in geoscience [25], rely on the idea that the dynamics of a given system may repeat to some extent. Given a set of previously observed or analysed data, one may retrieve examples similar to a current state in this set, such that the future of this current state may be forecasted from the known evolution of these similar situations. The lack of large-scale dataset along with the computational complexity of analog methods has long limited their applicability. In this context, we recently introduced the analog data assimilation (AnDA) and demonstrated its relevance for the reconstruction of complex dynamical systems for partial observations, including sea surface dynamics [26,27]. Here, as stated in the next section, we further explore and evaluate AnDA schemes for the reconstruction of cloud-free SST fields from satellite-derived measurements.

The remainder of the paper is organized as follows. Section 2 briefly reviews the related work on data assimilation and introduces the main contributions of this work. Section 3 presents the considered data and case-study region. Section 4 describes the proposed AnDA methods for the reconstruction of cloud-free SST fields. Section 5 presents experimental results. Section 6 further discusses our key contributions and future work.

2. Problem Statement and Related Work

Data assimilation is the classic framework for the reconstruction of sea surface geophysical fields from partial satellite observation series [20,28]. Two main categories of data assimilation methods may be distinguished: variational and statistical data assimilation. Variational methods rely on a continuous setting and states data assimilation as the minimization of a variational cost. Statistical methods involve state-space models [20,28]. They formulate data assimilation as the maximization or estimation of the posterior likelihood of the state series given an observation series. The state refers to the geophysical parameter of interest, here a cloud-free SST field at a given time. In this work, we focus on statistical data assimilation methods, which provides a greater flexibility to model state dynamics as well as the relationship between the state series and the observation series [20]. They also avoid determining the adjoint of the dynamic operator, which may be complex while reaching state-of-the-art reconstruction performance [29].

The state-space model typically comprises two key components:

- A dynamical model which states the time evolution of the state. Within a discrete statistical framework, it comes to define the likelihood of the state at a given time given the state at the previous time;
- An observation model which relates the state to the observation, here the cloud-free SST field to the SST observation with missing data.

Among the variety of algorithms proposed to solve for statistical data assimilation issues, Ensemble Kalman filters and smoothers (EnKF and EnKS) are particularly popular. They demonstrate

both good assimilation performance and a high modeling flexibility [20]. It may also be noted that the optimal interpolation can be regarded as a statistical assimilation model, where the dynamical prior involves a Gaussian distribution, such that an analytical and numerical solution can be derived [12,20]. EnKS and EnKF may provide relevant solutions to implement optimal interpolation schemes for high-dimensional fields. The definition of the dynamical prior is a critical aspect of such model-driven assimilation scheme. Regarding ocean dynamics, the balance between modeling complexity and uncertainty is particularly complex. Especially, simplified models such as advection-diffusion or QG (Quasi-Geostrophic) priors [21] may only be valid approximations for specific space-time regions.

These issues have motivated the development of data-driven frameworks as an alternative to the definition of model-driven dynamical priors. We may for instance cite EOF-based (Empirical Orthogonal Function) interpolation techniques [14,15], which state interpolation issues as a matrix completion problem and iterates successive projections onto an EOF basis under the constraint of the observed data. Such techniques have been proven relevant for the reconstruction of large-scale SST fields. They may however lack some mathematically-sound interpretation in terms of data assimilation issue. Interestingly, the combination of analog forecasting operators and classic statistical assimilation schemes has led to the introduction of novel data-driven schemes for data assimilation, referred to as Analog Data Assimilation (AnDA) [24,26]. Especially, our previous work [26] presents an application of AnDA to the space-time interpolation of SST fields using patch-based and EOF decompositions. (We use in this paper the term patch to refer to a subset of $K \times K$ pixels centered on given pixel. K is the width of the patch. The term patch is widely used in image processing with the emergence of patch-based image representations [30,31]. Patch-based representations [30–33] provide means to encode the spatial structure of the images while providing a simple and computationally-efficient framework.

Here, we further extend AnDA for the spatiotemporal interpolation of SST fields to improve both reconstruction performance and computational efficiency. This work involves three main contributions:

- the introduction of conditional analog forecasting operators with a view to explicitly accounting for dependencies between the state to be reconstructed and auxiliary variables. In [24,26], the considered analog forecasting operators implicitly assumed the high-resolution component dX to be independent on the low-resolution component $\bar{X}$. Both theoretical and statistical studies [34,35] advocate for considering inter-scale dependencies, which relate to the multi-scale characteristics of ocean turbulence [35,36]. We show here that analog strategies are highly flexible to consider such conditioning;
- the introduction of an analog forecasting operator embedding physically-sound priors. We further benefit from the flexibility of analog operators to exploit the synergy between SSH (Sea Surface Height) [35,37] and SST. We investigate locally-linear analog forecasting operators where SSH is used as a complementary regressor;
- the reduction of the computational complexity of AnDA using a clustering-based analog forecasting operator. To improve the scalability of the proposed methodology, we show that we can significantly reduce the computational complexity of the analog data assimilation with no impact on reconstruction performance.

We demonstrate the relevance of these contributions through numerical experiments for real cloudy SST patterns and evaluate the computational complexity of the proposed models and their parallelization performance for future large-scale case-studies.

3. Data and Study Area

With a view to evaluating the proposed analog assimilation methodology detailed in the next section, we use reference gap-free L4 SST time series from which we create SST datasets with missing data using real missing data masks. We consider two gap-free L4 SST products:

- OSTIA SST: the OSTIA product delivered daily by the UK Met Office [12] with a 0.05° × 0.05° spatial resolution (approx. 5 km). The OSTIA analysis combines satellite data provided by infrared sensors (AVHRR, AATSR, SEVIRI), microwave sensors (AMSR-E, TMI) and in situ data from drifting and moored buoys.
- MW SST: the microwave optimally-interpolated product distributed by REMSS (http://www.remss.com/measurements/sea-surface-temperature/oisst-description/). This product combines daily microwave satellite measurements (TMI, AMSR-E, AMSR2, WindSat sensors) for a 0.25° × 0.25° resolution.

From a spectral analysis of the SST fields, it may be noted that the MW SST dataset involves greater energy level for scales below 100km than OSTIA SST dataset. For both datasets, we consider SST time series from January 2007 to December 2015 (January 2008 to December 2015) in a region off South Africa (150 × 300 pixels) from 0°E to 7°E and 22.5°S to 60°S. This region comprises the Aghulas current and combines highly-dynamic areas and periods off South Africa and not as active areas in the northern part of the case-study region which is characterized by warmer waters. This region is also characterized by a significant variability of the cloud cover up to very high missing data rates (e.g., above 70%). These characteristics make this region a relevant and representative testbed for SST interpolation issues.

As real cloud masking time series, we consider the cloud masks associated with the METOP-AVHRR SST time series (Ocean and Sea Ice Satellite Application Facility (OSI SAF) (2016). GHRSST L3C global sub-skin Sea Surface Temperature from the Advanced Very High Resolution Radiometer (AVHRR) on Metop satellites (currently Metop-B) (GDS V2) produced by OSI SAF (GDS version 2). NOAA National Centers for Environmental Information. Dataset) METOP-AVHRR product is a high-resolution infrared sensor, which may involve very high missing data rates in the case-study area (see Figure 1 for an example of cloud mask pattern).

As detailed in the next section, the proposed analog data assimilation models may benefit from multi-source data. More particularly, in the considered case-study, we explore the extent to which SSH data may be useful to improve the interpolation of the SST. As gridded and interpolated SSH field, we consider daily SSH data with a 0.25° × 0.25° resolution distributed by the CMEMS (Copernicus Marine Environment Monitoring Service, marine.copernicus.eu).

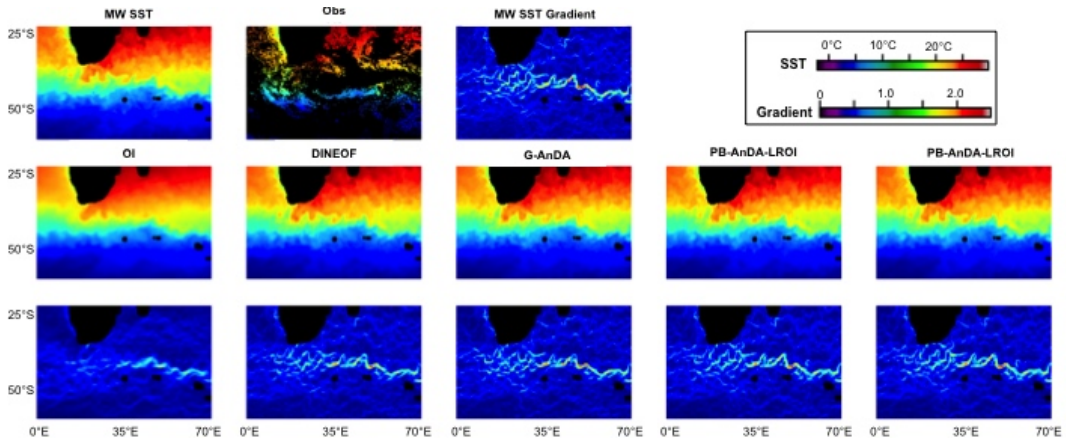

Figure 1. Reconstructed SST fields using OI, DINEOF, G-AnDA, PB-AnDA-LROI + dX + Z, PB-AnDA-LRM + dX + Z on day 150th for MW SST case-study: the first row depicts the reference SST field, the cloudy observation and the gradient magnitude; the second and third rows depict respectively the SST fields and their gradient magnitude for OI, DINEOF, G-AnDA, PB-AnDA-LROI + dX + Z, PB-AnDA-LRM + dX + Z. It may be noticed that PB-AnDA-LROI + dX + Z and PB-AnDA-LRM + dX + Z better reconstructs fine-scale structures for instance along the Aghulas return current as well as south of Madagascar island.

4. Method

4.1. Patch-Based Analog Data Assimilation

We consider the following scale-based decomposition of the SST fields:

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

$\bar{X}$ refers to a background field. It may be a given as a mean field as well as optimally-interpolated fields. dX refers to high-resolution component to be estimated.

Following [26], we consider an analog data assimilation for the high-resolution component dX. It involves a patch-based state-dependent dynamical operator. Let us denote by $\mathcal{P}_s$ the patch centered on grid site s and $X(\mathcal{P}_s, t)$ the patch-level state for field X on $\mathcal{P}_s$ at time t. The considered dynamical operator for $\mathcal{P}_s$ at time t is stated as

$$dX\left(\mathcal{P}_s, t\right) = \mathcal{M}_{X(\mathcal{P}_s, t-1)}\left(X(\mathcal{P}_p s, t-1), \eta(\mathcal{P}_s, t-1)\right) \tag{2}$$

where $\mathcal{M}_{X(\mathcal{P}_s, t-1)}$ is the state-dependent operator at time t for grid site s. η is a random perturbation. We further constrain this patch-based model through an EOF-based decomposition of each patch-level state $dX\left(\mathcal{P}_s, t\right)$.

$$dX\left(\mathcal{P}_s, t\right) = \sum_{n=1}^{N} \alpha_n(s,t) B_n \tag{3}$$

With B_n the nth principal component of the EOF and $\alpha_n(s,t)$ the associated EOF expansion coefficient for patch $\mathcal{P}_s$ at time t. N_{EOF} refers to the number of vectors of the EOF basis. B will denote the matrix formed by all principal components.

The state-dependent dynamical operator $\mathcal{M}_{X(\mathcal{P}_s, t-1)}$ is stated as a locally-linear analog forecasting operator [24,26]. We assume that we are provided with a reference dataset, referred to as catalog $\mathcal{C}$ which comprises pairs of states $\{X(\mathcal{P}_{s_i}, t_i - 1), X(\mathcal{P}_{s_i}, t_i)\}_i$ at two consecutive time steps, referred to respectively as analogs and successors. For a given kernel denoted by $\mathcal{K}$, let us denote by $a_k\left(s,t\right))$ the k^{th} analog (i.e., nearest-neighbor) of state $X(\mathcal{P}_s, t)$ in catalog $\mathcal{C}$ and $s_k\left(s,t\right))$ its successor. The locally-linear analog operator is stated as a multivariate linear regression in the EOF space between the analogs $\{a_k\left(s,t\right))\}_k$ and their successors $\{s_k\left(s,t\right))\}_k$ with a zero-mean Gaussian perturbation. The linear regression is fitted using a weighted least-square estimate with weights $\{\mathcal{K}(a_k\left(s,t\right), X(\mathcal{P}_s, t-1))\}_k$. The covariance of the Gaussian perturbation is estimated from the residual of the linear regression for the K pairs of analogs and successors. In [24,26], the regression variables are directly the states projected onto the EOF space. Here, as detailed in the next section, we consider different parameterization of the regression variables as well as of the kernel $\mathcal{K}$ to explore the potential conditioning of the dynamics of state dX by other variables (e.g., the low-resolution component $\bar{X}$ or a velocity field).

Given the observation model associated with the considered cloudy SST observations

$$Y(t,s) = X(t,s) + \epsilon(t,s),\ \forall s \in \Omega_t \tag{4}$$

With Ω_t the cloud-free region at time t, the reconstruction of high-resolution component dX given observation time series Y relies on the ensemble Kalman smoother (EnKS) associated with the considered analog dynamical operators. The EnKS is a forward-backward sequential algorithm. It represents the state at each time step from the mean and covariance of a set of members, which are evolved in time based on the analog forecasting operator and updated at each time step from the available observations using a Kalman-based recursion. We refer the reader to [24] for the details of the analog EnKS. Here, the analog EnKS is applied independently to overlapping patch locations. We then reconstruct field dX as a mean over overlapping patches. An additional postprocessing step is

applied to remove possible patch-related blocky artifacts using a patch-based and global EOF filtering. The resulting workflow is sketched in Figure 2.

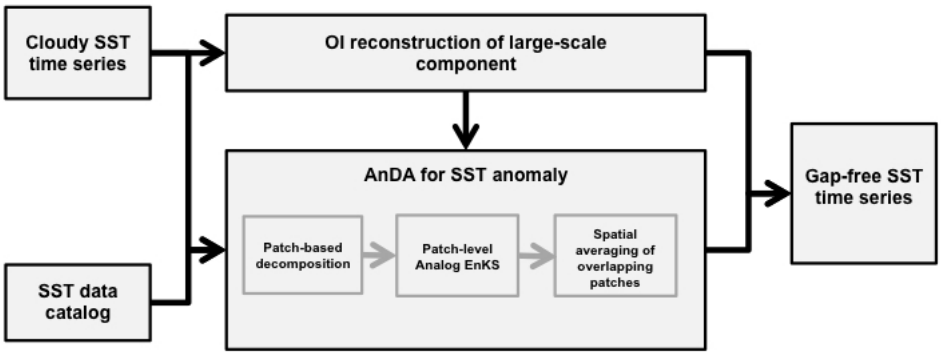

Figure 2. Workflow of the proposed framework for the reconstruction of gap-free SST time series from cloudy SST data: given a cloudy SST field time series, it first applies an optimal interpolation to reconstruct the large-scale component $\widehat{X}$ and second the analog data assimilation (AnDA) of the anomaly dX (cf. (Equation (1))). This second step exploits a reference SST catalog and is constrained by the reconstructed large-scale component. The resulting gap-free SST time series is the sum of the large-scale component $\widehat{X}$ and of the anomaly dX. We also sketch the main steps involved in the AnDA scheme.

4.2. Conditional and Physically-Derived Analog Forecasting Operators

Let us denote by U a co-variable with the same space-time resolution as field X. The Conditioning of analog forecasting operator $\mathcal{M}_{X(\mathcal{P}_s,t-1)}$ at time t and patch $\mathcal{P}_s$ may be issued:

- from the selection of analogs based on both variables dX and U, and not solely based on dX as in [24,26]. This comes to take into account variable Z in kernel $\mathcal{K}$. We typically consider a parameterization of kernel $\mathcal{K}$ as $\mathcal{K}_{dX} \cdot \mathcal{K}_U$ using kernels applied respectively to fields dX and Z. Here, we will consider a Gaussian kernel for $\mathcal{K}_{dX}$ and a correlation-based kernel for kernel $\mathcal{K}_U$. It may be noted that the considered kernels only exploit the spatial dimensions;
- from the fit of a multivariate linear regression using both dX and U, or transformed version of U, as regression variables and not solely based on dX as in [24,26]. For instance, following previous studies [34,38], one may consider the low-resolution field $\bar{X}$ as a potentially-relevant information to improve the forecasting of the high-resolution field dX.

It may be emphasized that a given co-variable may be used only for one of these two types of conditioning.

We also explore the potential relationship between locally-linear analog forecasting operator and physical operator. As a sea surface geophysical tracer, advection-diffusion priors may be regarded as relevant first-order approximations [21]. The advection-diffusion prior is given by

$$\partial_t X + \langle \omega, \nabla X \rangle = \kappa \Delta X \tag{5}$$

With ω the sea surface velocity field and κ the diffusion coefficient. Given that satellite-derived altimeter fields, denoted here by Z, provide a low-resolution estimate of the sea surface velocity fields, field ω may be written as $\nabla^{\perp} Z + \delta\omega$ with $\delta\omega$ the unresolved velocity component. Using decomposition Equation (1), advection-diffusion prior Equation (5) suggests considering a locally-linear patch-based analog forecasting operator Equation (2) where both variables dX, $\bar{X}$ and Z are considered as regression variables. Similarly to EOF decomposition Equation (3) for field dX, we also consider patch-based EOF decompositions for fields $\bar{X}$ and Z to constrain the estimation of the analog locally-linear operator.

It may be noted that this inference of such multi-modal analog forecasting operators relate to the approximation of the underlying unresolved velocities from local analogs.

4.3. Computationally-Efficient Analog Assimilation Strategies

An important goal of this study is to evaluate the computational complexity of the proposed analog assimilation strategies and their parallelization properties. By construction, the considered patch-level decomposition relies on the independent processing of each patch location for the considered grid. This ensures the computational complexity to evolve linearly with the number of grid points. The considered EnKS procedure involves two main steps: the forecasting and the analysis step. Given the relatively low-dimensional EOF-based representation of each patch, the computational complexity of the analog EnKS mainly relates to the ananlog forecasting step. In the standard version used in [24], the computational complexity may be decomposed as $N_M \cdot (C_{search} + C_{fit} + C_{forecasting})$ with N_M the number of members used to represent the state at each time step, C_{search} the computational cost of the search for analogs, C_{fit} the computational cost of the fit of the analog forecasting operator and $C_{forecasting}$ the computational cost of the application of the fitted forecasting operator. The first two ones are obviously the most important ones.

To speed up the search for local analogs, we can benefit from large research effort dedicated to nearest-neighbor search, especially approximate nearest-neighbor search [39,40]. Here, we consider FLANN (Fast Library for Approximate Nearest Neighbors) frameworkavailable at http://www.cs.ubc.ca/research/flannforadditionaldetails), which is among the state-of-the-art schemes for approximate nearest-neighbor search. It relies on an offline computation of a tree-based indexing structure. We let the reader [40,41] and FLANN (Fast Library for Approximate Nearest Neighbors) library.

Importantly, it may be noted that at a given time step many members can be expected to share similar dynamics. Therefore, fitting a local analog forecasting operator for each member as in [24,26] is expected to be computationally-redundant. To reduce this computational redundancy, we introduce a clustering-based strategy. For a given time step, we constrain the computational complexity to a given number of analog forecasting operator fit, denoted by N_{Fit}. We first clusterized the members into N_{Fit} using a K-means procedure [42]. We then fit an analog forecasting model for each cluster using the analogs and successors to the center of the cluster. For the forecasting step, we apply to each member the analog forecasting operator of the cluster it is assigned to. Overall, this clustering-based strategy leads to cost $N_{Fit} \cdot C_{fit}$ to compare to the original $N_M \cdot C_{fit}$. Here, we typically set N_{Fit} to 3 whereas $N_M = 100$.

4.4. Experimental Setting

Computational setting: The considered experiments, especially regarding the evaluation of the computational complexity of the proposed methods, have been implemented onto Teralab platform (https://www.teralab-datascience.fr/fr/) using a virtual machine with the following setting: 30 CPUs with a 64 G RAM (24 CPUs are used for processing tasks and others as backup or for background tasks). All experiments were run using Python and PB-AnDA Python library available at https://github.com/rfablet/PB_ANDA. We used Multiprocessing Python module to implement AnDA onto the considered multi-core platform.

Benchmarked models and algorithms: We consider different patch-based analog assimilation models, referred to as PB-AnDA, corresponding to different parameterizations of the analog forecasting operators:

- for the low-resolution component $\bar{X}$, we consider two options: (i) optimally-interpolated fields projected onto a region-level EOF decomposition with 20 components which resolve spatial scales up to approximately 100 km, (ii) the mean field. The first one is referred to LROI and second one to LRM;
- for the search for analogs, we explored both a simple kernel with no conditioning by the low-resolution component, such that $\mathcal{K} = \mathcal{K}_{dX}$ and a kernel $\mathcal{K} = \mathcal{K}_{dX} * \mathcal{K}_Z$ with $Z = \|\nabla \bar{X}\|$ to

introduce a conditioning of the analog forecasting operators by the low-resolution gradient magnitude as suggested in [34]. As both settings resulted in very similar interpolation performance (e.g., RMSE of 0.24 for MW SST dataset for both settings), we only report results for the simplest kernel choice (i.e., $\mathcal{K} = \mathcal{K}_{dX}$) in the subsequent analysis.

- three types of regression variables were evaluated: locally-linear operators using only dX as regression variables (dX), using dX and $\bar{X}$ as regression variables ($dX + \bar{X}$) and using dX and Z as regression variables ($dX + Z$). A fully-developed locally-linear approximation of an advection-diffusion prior would consist in considering both dX, $\bar{X}$ and Z as regression variables. It resulted in the same performance as considering only dX and Z (see Table 1) and was not included in the reported results. We might recall that all locally-linear models are fitted within EOF subspaces.

Table 1. Interpolation performance of PB-AnDA models for the MW SST case-study for three zones of interest in the case-study region: we refer the reader to the main text for the description of the different PB-AnDA parameterizations.

PB-AnDA	LROI			LRM		
	dX	$dX + Z$	$dX + \bar{X}$	dX	$dX + Z$	$dX + \bar{X}$
Zone 1	0.35 ± 0.06	$\mathbf{0.34 \pm 0.05}$	0.35 ± 0.06	0.34 ± 0.06	$\mathbf{0.33 \pm 0.05}$	0.34 ± 0.06
Zone 2	0.33 ± 0.09	$\mathbf{0.32 \pm 0.08}$	0.33 ± 0.09	0.32 ± 0.08	$\mathbf{0.30 \pm 0.07}$	0.32 ± 0.08
Zone 3	$\mathbf{0.18 \pm 0.04}$	$\mathbf{0.18 \pm 0.04}$	$\mathbf{0.18 \pm 0.04}$	0.19 ± 0.04	0.19 ± 0.04	0.19 ± 0.04

Overall, we refer to a specific model as follows. For instance, model $PB\text{-}AnDA + LROI + dX + \bar{X} + Z$ implements the patch-based analog assimilation with: (i) optimally-interpolated fields as low-resolution component and (ii) locally-linear analog forecasting operators with auxiliary variables $\bar{X}$ and Z. All patch-based analog assimilation models involve similar parameter setting regarding the number of members, $N_M = 100$, and the patch-based EOF decomposition with $N_{EOF} = 50$, which account for 96% of the total variance of the SST datasets.

For benchmarking purposes, we considered two state-of-the-art interpolation techniques and a global AnDA model:

- a classic optimal interpolation with a Gaussian space-time covariance structure: the spatial and time correlation lengths were tuned from cross-validation experiments for the considered SST datasets to respectively 3 days and 100 km. This interpolation is referred to as OI and implemented using [43];
- a DINEOF interpolation [14]: the EOF-based interpolation comes to iteratively project the reconstructed field onto the EOF basis while modifying only SST values for missing data areas. We use 40 EOF components to account for about 95% of the total variance. This interpolation referred to as DINOEF is applied globally onto the entire case-study region.
- a direct application of AnDA over the entire region: this interpolation referred to as G-AnDA exploits the same EOF decomposition as DINEOF and $N_M = 100$ members in the implemented AnDA.

5. Results

5.1. Interpolation Performance

We first compare interpolation performance of the considered methods, namely OI, DINEOF, G-AnDA and PB-AnDA + LROI + dX for the both OSTIA and MW case-studies (Tables 2 and 3). Similar conclusions can be drawn from these experiments with a significant gain of the proposed patch-based AnDA model compared to the three other approaches. For instance, We report a relative gain up to 50% in RMSE w.r.t. OI and of 40% w.r.t. DINEOF. Though the direct application of AnDA to the entire region lead to a slight improvement (e.g., mean RMSE of 0.38 for G-AnDA w.r.t. 0.40

for DINEOF and 0.43 for OI on MW SST dataset), the additional relative gain greater than 35% in RMSE of the patch-based version PB-AnDA + LROI + dX emphasizes the relevance of the proposed multi-scale and patch-based decomposition to account for fine-scale structures. The analysis of the mean correlation coefficients between the interpolated fields and the reference fields for scales below 100 km leads to the same conclusion.

Table 2. Interpolation performance for the MW SST case-study: mean root mean square error (RMSE) and correlation coefficients with the MW SST for OI, DINEOF, G-AnDA and PB-AnDA methods. We refer the reader to the main text for the details on the considered parameterizations.

Criterion	RMSE	Correlation
OI	0.48 ± 0.05	0.69 ± 0.07
DINEOF	0.40 ± 0.04	0.79 ± 0.04
G-AnDA	0.38 ± 0.04	0.81 ± 0.03
PB-AnDA + LROI + dX	0.24 ± 0.03	0.93 ± 0.02

Table 3. Interpolation performance for the OSTIA SST case-study: mean root mean square error (RMSE) and correlation coefficients with the OSTIA SST for OI, DINEOF, G-AnDA and PS-MS-AnDA methods. We refer the reader to the main text for the details on the considered parameterizations.

Criterion	RMSE	Correlation
OI	0.42 ± 0.11	0.83 ± 0.07
DINEOF	0.40 ± 0.10	0.86 ± 0.06
G-AnDA	0.38 ± 0.08	0.87 ± 0.04
PB-AnDA + LROI + dX	0.22 ± 0.04	0.90 ± 0.03

Based on the above results, we further compared the performance of the different parameterization of the proposed Pb-AnDA models. The higher energy level of MW SST fields for scales below 100 km made the MW SST case-study more appropriate for this analysis. We evaluate the interpolation performance for PB-AnDA modes using respectively dX, $dX + Z$ and $dX + \bar{X}$ variables using both the optimally-interpolated field (LROI) and the yearly mean (LRM) as low-resolution background. We report in Table 1 the RMSE for three specific zones: a first zone from (10°E, 36.25°S) to (56.25°E, 45°S), a second zone from (55°E, 38.75°S) to (75°E, 47.5°S) and a third zone from (35°E, 26.25°S) to (55°E, 35°S). The first two zones depict highly-dynamical patterns, whereas the dynamics on the third one are not as intense. Whereas auxiliary variables do not bring any improvement for Zone 3 for both LROI and LRM settings, a slight mean improvement is reported when considering Z for the two other zones (i.e., the EOF-based decomposition of the SSH field) as auxiliary variable (e.g., RMSE values of 0.32 vs. 0.30 for PB-AnDA-LRM-dX and PB-AnDA-LRM-dX + Z in Zone 2). Surprisingly, the exploitation of the optimally-interpolated background (LROI setting) may be outperformed by LRM setting (e.g., RMSE of 0.32 for PB-AnDA-LROI-dX and 0.3 for PB-AnDA-LRM-dX in Zone 2). This may suggest that the space-time smoothing of the optimal interpolation for highly-dynamical situations may result in local biases. To check for this hypothesis, we run a complementary experiment using 5-daily SST field in order to simulate even higher-dynamical situations. As reported in Table 4, these experiments further pinpoint the relevance of PB-AnDA-LRM-dX + Z setting to deal with highly-dynamical situations. Example reported in Figures 1 and 3 illustrates these conclusions. Visually, the use of the SSH as auxiliary variable (Z) leads to a better reconstruction of the fine-scale structures. This is further illustrated in Figure 4 for Zone 2.

As a synthesis, we report in Figure 5 the time series of the mean RMSE and correlation for the MW SST case-study for OI (blue), DINEOF, G-AnDA, PB-AnDA-LROI-dX + Z, PB-AnDA-LRM-dX + Z. These results clearly emphasize the relevance of PB-AnDA models. Besides lower RMSE values and higher correlation coefficients, PB-AnDA models also depict a much lower variability. Similarly to the zone-specific results discussed above, LRM and LROI settings lead to a very similar performance (mean RMSE of 0.239 for PB-AnDA-LRM-dX + Z and of 0.241 for PB-AnDA-LROI-dX + Z). Though this

may not appear significant when considering a spatiotemporal mean, this is the case when considering specific dates and zones corresponding to highly-dynamical situations. I may be noted that the overall computational complexity of PB-AnDA-LRM-$dX + Z$ is significantly lower as it did not require the computation of the OI field as low-resolution background.

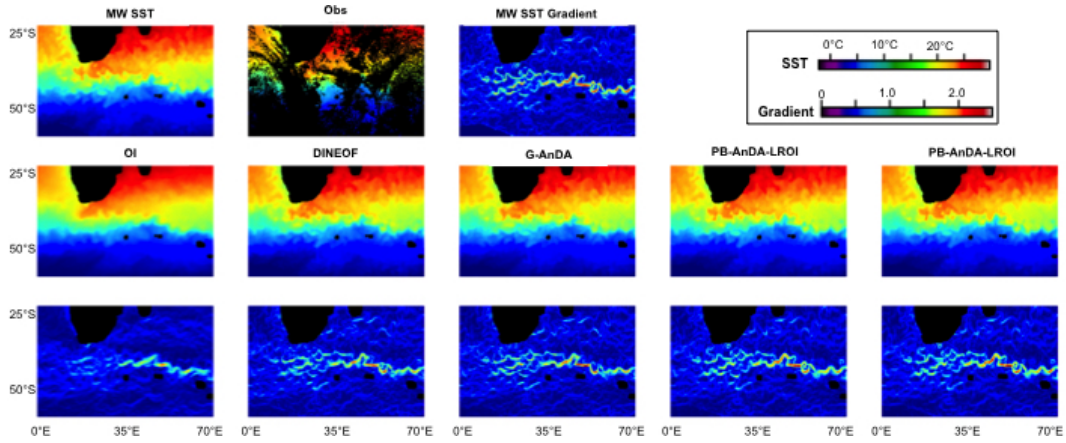

Figure 3. Reconstructed SST fields using OI, DINEOF, G-AnDA, PB-AnDA-LROI + $dX + Z$, PB-AnDA-LRM + $dX + Z$ on day 150 for MW SST case-study: refer to Figure 1. It may be noticed that only PB-AnDA-LROI + $dX + Z$ and PB-AnDA-LRM + $dX + Z$ retrieves the fine-scale eddy-like structure off South Africa on this particular date.

Table 4. Interpolation performance of PB-AnDA models for a 5-daily MW SST case-study for three zones of interest in the case-study region: we refer the reader to the main text for the description of the different PB-AnDA parameterizations.

PB-AnDA	LROI			LRM		
	dX	$dX + Z$	$dX + \bar{X}$	dX	$dX + Z$	$dX + \bar{X}$
Zone 1	0.49 ± 0.10	$\mathbf{0.47 \pm 0.10}$	0.49 ± 0.10	0.51 ± 0.12	$\mathbf{0.45 \pm 0.09}$	0.51 ± 0.12
Zone 2	0.48 ± 0.14	$\mathbf{0.43 \pm 0.12}$	0.48 ± 0.14	0.49 ± 0.17	$\mathbf{0.38 \pm 0.10}$	0.48 ± 0.17
Zone 3	0.23 ± 0.06	$\mathbf{0.22 \pm 0.06}$	0.23 ± 0.06	0.23 ± 0.06	$\mathbf{0.22 \pm 0.06}$	0.23 ± 0.06

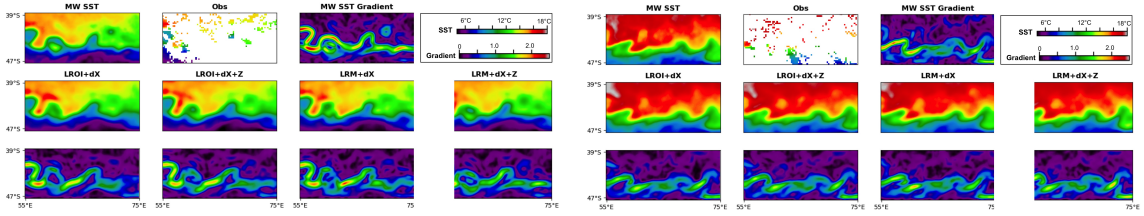

Figure 4. Comparison of assimilation results for Zone 2 and case-study MW SST for different parameterization of the PB-AnDA models: we report the example for two dates, respectively the 104th and 309th of the MW SST time series, in panels (**a**,**b**). For each panel, the first row depicts the MW SST field (MW SST), the cloudy SST observation (Obs) and the gradient field (MW SST Gradient). The second and third displays respectively the reconstructed SST fields and their gradient magnitude for Pb-AnDA using LROI + dX, $dX + Z$, LROI + LRM + dX, LRM + $dX + Z$ parameterizations.

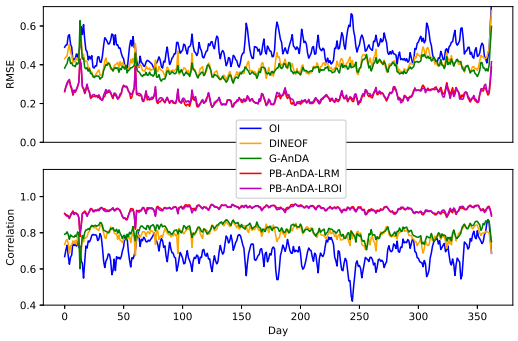

Figure 5. Mean RMSE and correlation time series for MW SST case-study: OI (blue), DINEOF (orange), G-AnDA (green), PB-MS-AnDA-LROI (red), PB-MS-AnDA-LRM (purple) methods.

5.2. Computational Complexity and Scalability

Regarding computational complexity issues, we evaluate the computational time of the PB-AnDA models with respect to the number processing cores. Figure 6 emphasizes the scalability of PB-AnDA models with good parallelization performance, as the computational time almost reaches the optimal linear decrease w.r.t. the number of cores in logarithmic scale. This parallelization performance directly relates to the proposed patch-based setting which leads to the independent sequential assimilation of the considered patches.

These experiments also stress the significant reduction of the computational complexity resulting from the clustering-based analog forecasting operators. When considering a 12-core architecture, the computational time is for instance reduced by a factor of about 4 between the proposed clustering-based scheme compared with the original one [24].

Overall, for the considered case-study region with 194 patches and the considered multi-core architecture, the overall computation time required by PB-AnDA is significantly less than that of G-AnDA (23 min vs. 82 min), though higher than that of DINEOF (23 min vs. 4 min). These results support an operational application of the proposed AnDA models on a global scale for high-resolution SST fields using state-of-the-art multi-core architecture.

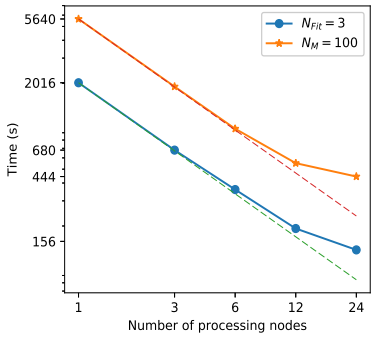

Figure 6. Parallelization performance of the proposed PB-AnDA setting: we report in logarithmic scale the computational time of the assimilation of 24 patches using the standard PB-AnDA with $N_M = 100$ members (orange,-) and using the clustering-based version with $N_{Fit} = 3$ clusters (blue,-). The dashed lines indicate the computational time of a theoretically-optimal multi-core parallelization.

6. Conclusions

We presented an application of the analog data assimilation [24] to the interpolation of SST fields. Using patch-based and EOF-based decompositions as in [26], the main contributions of this study

are three-fold: (i) the introduction of conditional and physically-driven analog forecasting operators, (ii) the reduction of the computational complexity of PB-AnDA models with clustering-based analog forecasting operators, (iii) the demonstration of the scalability of PB-AnDA models to scale up to large scale-datasets. Overall, this study supports the investigation of the operational application of PB-AnDA models for an improved spatio-temporal interpolation of SST fields compared with optimal interpolation [12] and EOF-based schemes [14]. Among the issues to be delt with, the size and nature of the SST catalogs to be archived is certainly a critical question. Future work should further explore the extent to which purely-observation-based catalog may be self-sufficient or appropriately complemented by numerical simulation datasets. Preliminary results suggest that purely-observation-based catalogs might be a relevant option. Future should also investigate how AnDA may also provide a flexible framework to combine multi-source and multi-scale SST data through adapted observation models [10].

Beyond the reconstruction of gapfilled SST fields, we believe that the reported experiments illustrate the potential of PB-AnDA models for the reconstruction of geophysical products from remote sensing data, especially other sea surface tracers such as SSH (Sea Surface Height), SSS (Sea Surface Salinity) and ocean colour, as well as atmospheric variables. It might be noted that our recent application on the interpolation of altimeter-derived SSH fields further supports this potential [27]. For such applications, the relevance of PB-AnDA models is expected to strongly depend on one hand on the availability of large-scale simulation or observation-driven datasets to build representative catalogs of exemplars of the range of space-time scales of interests, and, on the other hand, on the validity of the assumption that state dynamics are locally-linear with respect to the considered regressors.

Acknowledgments: This work was supported by ANR (Agence Nationale de la Recherche, grant ANR-13-MONU-0014), Labex Cominlabs (grant SEACS), Teralab (grant TIAMSEA) and CNES (grant OSTST-MANATEE).

Author Contributions: R.F., R.L. and B.C. conceived and designed the experiments; P.V. performed the experiments; R.F. P.V. and R.L. analyzed the data; P.V., R.L. and PH.H. contributed reagents/materials/analysis tools; R.F. wrote the paper." Authorship must be limited to those who have contributed substantially to the work reported.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Donlon, C.J.; Minnett, P.J.; Gentemann, C.; Nightingale, T.J.; Barton, I.J.; Ward, B.; Murray, M.J. Toward Improved Validation of Satellite Sea Surface Skin Temperature Measurements for Climate Research. *J. Clim.* **2002**, *15*, 353–369.
2. Merchant, C.J.; Embury, O.; Roberts-Jones, J.; Fiedler, E.; Bulgin, C.; Corlett, G.K.; Good, S.; McLaren, A.; Rayner, N.; Morak-Bozzo, S.; et al. Sea surface temperature datasets for climate applications from Phase 1 of the European Space Agency Climate Change Initiative (SST CCI). *Geosci. Data J.* **2014**, *1*, 179–191.
3. Hollmann, R.; Merchant, C.J.; Saunders, R.; Downy, C.; Buchwitz, M.; Cazenave, A.; Chuvieco, E.; Defourny, P.; de Leeuw, G.; Forsberg, R.; et al. The ESA Climate Change Initiative: Satellite Data Records for Essential Climate Variables. *Bull. Am. Meteorol. Soc.* **2013**, *94*, 1541–1552.
4. Filipponi, F.; Valentini, E.; Taramelli, A. Sea Surface Temperature changes analysis, an Essential Climate Variable for Ecosystem Services provisioning. In Proceedings of the 9th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp), Brugge, Belgium, 27–29 June 2017; pp. 1–8.
5. Li, J.; Heap, A.D. Spatial interpolation methods applied in the environmental sciences: A review. *Environ. Model. Softw.* **2014**, *53*, 173–189.
6. Li, J.; Wang, P.; Han, H.; LI, J.; Zhang, J. On the assimilation of satellite sounder data in cloudy skies in numerical weather prediction models. *J. Meteorol. Res.* **2016**, *30*, 169–182.
7. Penny, S.G.; Hamill, T.M. Coupled Data Assimilation for Integrated Earth System Analysis and Prediction. *Bull. Am. Meteorol. Soc.* **2017**, *98*, ES169–ES172.
8. Waters, J.; Lea, D.J.; Martin, M.J.; Mirouze, I.; Weaver, A.; While, J. Implementing a variational data assimilation system in an operational 1/4 degree global ocean model. *Q. J. R. Meteorol. Soc.* **2015**, *141*, 333–349.

9. Wada, A.; Kunii, M. The role of ocean-atmosphere interaction in Typhoon Sinlaku (2008) using a regional coupled data assimilation system. *J. Geophys. Res. Oceans* **2017**, *122*, 3675–3695.
10. Guan, L.; Kawamura, H. Merging Satellite Infrared and Microwave SSTs: Methodology and Evaluation of the New SST. *J. Oceanogr.* **2004**, *60*, 905–912.
11. Wentz, F.; Gentemann, C.; Smith, D.; Chelton, D. Satellite Measurements of Sea Surface Temperature through Clouds. *Science* **2000**, *288*, 847–850.
12. Donlon, C.J.; Martin, M.; Stark, J.; Roberts-Jones, J.; Fiedler, E.; Wimmer, W. The Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system. *Remote Sens. Environ.* **2012**, *116*, 140–158.
13. Donlon, C. The Next Generation of Multi-Sensor Merged Sea Surface Temperature Data Sets for Europe. In *Remote Sensing of the European Seas*; Barale, V., Gade, M., Eds.; Springer: Berlin, Germany, 2008; pp. 177–188, doi:10.1007/978-1-4020-6772-3_14.
14. 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.
15. Beckers, J.M.; Rixen, M. EOF Calculations and Data Filling from Incomplete Oceanographic Datasets. *J. Atmos. Ocean. Technol.* **2003**, *20*, 1839–1856.
16. Reynolds, R.W.; Smith, T.M. Improved Global Sea Surface Temperature Analyses Using Optimum Interpolation. *J. Clim.* **1994**, *7*, 929–948.
17. Reynolds, R.W.; Smith, T.M.; Liu, C.; Chelton, D.B.; Casey, K.S.; Schlax, M.G. Daily High-Resolution-Blended Analyses for Sea Surface Temperature. *J. Clim.* **2007**, *20*, 5473–5496.
18. Tandeo, P.; Ailliot, P.; Autret, E. Linear Gaussian state-space model with irregular sampling: Application to sea surface temperature. *Stoch. Environ. Res. Risk Assess.* **2010**, *25*, 793–804.
19. Dash, P.; Ignatov, A.; Martin, M.; Donlon, C.; Brasnett, B.; Reynolds, R.W.; Banzon, V.; Beggs, H.; Cayula, J.F.; Chao, Y. Group for High Resolution Sea Surface Temperature (GHRSST) analysis fields inter-comparisons—Part 2: Near real time web-based level 4 SST Quality Monitor (L4-SQUAM). *Deep Sea Res. Part II Top. Stud. Oceanogr.* **2012**, *77*, 31–43.
20. Evensen, G. *Data Assimilation*; Springer: Berlin/Heidelberg, Germany, 2009.
21. 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.
22. Sirjacobs, D.; Alvera-Azcárate, A.; Barth, A.; Lacroix, G.; Park, Y.; Nechad, B.; Ruddick, K.; Beckers, J.M. Cloud filling of ocean colour and sea surface temperature remote sensing products over the Southern North Sea by the Data Interpolating Empirical Orthogonal Functions methodology. *J. Sea Res.* **2011**, *65*, 114–130.
23. Zhao, Z.; Giannakis, D. Analog Forecasting with Dynamics-Adapted Kernels. *arXiv* **2014**, arXiv:physics1412.3831.
24. Lguensat, R.; Tandeo, P.; Aillot, P.; Fablet, R. The Analog Data Assimilation. *Mon. Weather Rev.* **2017**, doi:10.1175/MWR-D-16-0441.1.
25. Lorenz, E.N. Deterministic Nonperiodic Flow. *J. Atmos. Sci.* **1963**, *20*, 130–141.
26. Fablet, R.; Viet, P.H.; Lguensat, R. Data-Driven Models for the Spatio-Temporal Interpolation of Satellite-Derived SST Fields. *IEEE Trans. Comput. Imag.* **2017**, *3*, 647–657.
27. Lguensat, R.; Huynh Viet, P.; Sun, M.; Chen, G.; Fenglin, T.; Chapron, B.; Fablet, R. *Data-Driven Interpolation of Sea Level Anomalies Using Analog Data Assimilation*. Available online: https://hal.archives-ouvertes.fr/hal-01609851 (accessed on 4 October 2017).
28. Asch, M.; Bocquet, M.; Nodet, M. *Data Assimilation*; Fundamentals of Algorithms, Society for Industrial and Applied Mathematics: Heidelberg, Germany, 2016; doi:10.1137/1.9781611974546.
29. Fairbairn, D.; Pring, S.R.; Lorenc, A.C.; Roulstone, I. A comparison of 4DVar with ensemble data assimilation methods. *Q. J. R. Meteorol. Soc.* **2014**, *140*, 281–294.
30. Buades, A.; Coll, B.; Morel, J.M. A non-local algorithm for image denoising. In Proceedings of theCVPR'05 IEEE Conference on Computer Vision and Pattern Recognition, San Diego, USA, 20–25 June 2005; Volume 2, pp. 60–65.
31. Mairal, J.; Bach, F.; Ponce, J. Sparse Modeling for Image and Vision Processing. *Found. Trends Comput. Graph. Vis.* **2014**, *8*, 85–283.
32. Freeman, W.T.; Liu, C. Markov Random Fields for Super-Resolution. In *Advances in Markov Random Fields for Vision and Image Processing*; Blake, A., Kohli, P., Rother, C., Eds.; MIT Press: Cambridge, MA, USA, 2011.

33. Criminisi, A.; Perez, P.; Toyama, K. Region filling and object removal by exemplar-based image inpainting. *IEEE Trans. Image Process.* **2004**, *13*, 1200–1212.
34. Fablet, R.; Rousseau, F. Joint Interpolation of Multisensor Sea Surface Temperature Fields Using Nonlocal and Statistical Priors. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 2665–2675.
35. Klein, P.; Isern-Fontanet, J.; Lapeyre, G.; Roullet, 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.
36. Bernard, D.; Boffetta, G.; Celani, A.; Falkovich, G. Inverse Turbulent Cascades and Conformally Invariant Curves. *Phys. Rev. Lett.* **2007**, *98*, 024501.
37. Tandeo, P.; Ailliot, P.; Ruiz, J.; Hannart, A.; Chapron, B.; Cuzol, A.; Monbet, V.; Easton, R.; Fablet, R. Combining Analog Method and Ensemble Data Assimilation: Application to the Lorenz-63 Chaotic System. In *Machine Learning and Data Mining Approaches to Climate Science*; Lakshmanan, V., Gilleland, E., McGovern, A., Tingley, M., Eds.; Springer: Berline, Germany, 2015; pp. 3–12.
38. 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.
39. Iwamura, M.; Sato, T.; Kise, K. What Is the Most Efficient Way to Select Nearest Neighbor Candidates for Fast Approximate Nearest Neighbor Search. In Proceedings of 2013 IEEE International Conference on Computer Vision (ICCV), Sydney, NSW, Australia, 1-8 December 2013; pp. 3535–3542.
40. Muja, M.; Lowe, D.G. Scalable Nearest Neighbor Algorithms for High Dimensional Data. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 2227–2240.
41. Muja, M.; Lowe, D.G. Fast approximate nearest neighbors with automatic algorithm configuration. In Proceedings of the VISAPP'09 International Conference on Computer Vision Theory and Applications, Lisboa, Portugal, 5–8 February 2009; Volume 2, p. 2.
42. Duda, R.O.; Hart, P.; Stork, D. *Pattern Classification*; John Wiley & Sons: New York, NY, USA, 2012.
43. 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.