



**HAL**  
open science

# Neural variational Data Assimilation with Uncertainty Quantification using SPDE priors

Maxime Beauchamp, Ronan Fablet, Simon Benaichouche, Pierre Tandeo,  
Nicolas Desassis, Bertrand Chapron

► **To cite this version:**

Maxime Beauchamp, Ronan Fablet, Simon Benaichouche, Pierre Tandeo, Nicolas Desassis, et al..  
Neural variational Data Assimilation with Uncertainty Quantification using SPDE priors. 2024. hal-  
04675468

**HAL Id: hal-04675468**

**<https://imt.hal.science/hal-04675468v1>**

Preprint submitted on 22 Aug 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

1 **Neural variational Data Assimilation with Uncertainty Quantification using**  
2 **SPDE priors**

3 Maxime Beauchamp,<sup>a,b</sup> Ronan Fablet,<sup>a</sup> Simon Benaichouche,<sup>a</sup> Pierre Tandeo,<sup>a</sup> Nicolas  
4 Desassis,<sup>c</sup> Bertrand Chapron,<sup>d</sup>

5 <sup>a</sup> *IMT Atlantique, 655 Av. du Technopôle, 29280 Plouzané, France*

6 <sup>b</sup> *Danish Meteorological Institute, Lyngbyvej 100, 2100 Copenhagen, Denmark*

7 <sup>c</sup> *Mines ParisTech, Centre de Géosciences, 35 Rue Saint-Honoré, 77300 Fontainebleau, France*

8 <sup>d</sup> *IFREMER, 1625 Rte de Sainte-Anne, 29280 Plouzané, France*

9 *Corresponding author: Maxime Beauchamp, maxime.beauchamp@imt-atlantique.fr*

10 *Simon Benaichouche's current affiliation: INRIA, Rennes, France*

11 ABSTRACT: The spatio-temporal interpolation of large geophysical datasets has historically been  
12 addressed by Optimal Interpolation (OI) and more sophisticated equation-based or data-driven  
13 Data Assimilation (DA) techniques. Recent advances in the deep learning community enables  
14 to address the interpolation problem through a neural architecture incorporating a variational  
15 data assimilation framework. The reconstruction task is seen as a joint learning problem of  
16 the prior involved in the variational inner cost, seen here as a projection operator of the state,  
17 and the gradient-based minimization of the latter. Both prior models and solvers are stated  
18 as neural networks with automatic differentiation which can be trained by minimizing a loss  
19 function, typically the mean squared error between some ground truth and the reconstruction.  
20 Such a strategy turns out to be very efficient to improve the mean state estimation, but still needs  
21 complementary developments to quantify its related uncertainty. In this work, we use the theory  
22 of Stochastic Partial Differential Equations (SPDE) and Gaussian Processes (GP) to estimate both  
23 space-and time-varying covariance of the state. Our neural variational scheme is modified to  
24 embed an augmented state formulation with both state and SPDE parametrization to estimate. We  
25 demonstrate the potential of the proposed framework on a spatio-temporal GP driven by diffusion-  
26 based anisotropies and on realistic Sea Surface Height (SSH) datasets. We show how our solution  
27 reaches the OI baseline in the Gaussian case. For nonlinear dynamics, as almost always stated  
28 in DA, our solution outperforms OI, while allowing for fast and interpretable online parameter  
29 estimation.

## 30 **1. Introduction**

31 Over the last decade, the emergence of large spatio-temporal datasets both coming from remote  
32 sensing satellites and equation-based numerical simulations has been noticed in Geosciences. As  
33 a consequence, the need for statistical methods able to handle both the size and the underlying  
34 physics of these data is growing. Data assimilation (DA) is the traditional framework used by  
35 geoscientists to merge these two types of information, data and model, by propagating information  
36 from observational data to areas with missing data. When no deterministic numerical outputs  
37 are available, the classic approach stems from the family of the so-called Optimal Interpolation  
38 (OI) techniques, also being at the core of the statistical DA methods (Asch et al. 2016). For the  
39 production of gridded geophysical maps, most of operational products either derived from OI  
40 and DA schemes allow us to estimate the mesoscale components of the targeted variables. In  
41 the specific case of Sea Surface Height (SSH) for instance, even if satellite altimetry provides  
42 capabilities to inform the mesoscale ocean geostrophic currents, the gridded product fails in  
43 merging observations and background model with consistent temporal and spatial resolutions able  
44 to retrieve fine mesoscale structures lower than 150-200 km at mid-latitudes, while they are key  
45 for general ocean circulation (Su et al. 2018). Then, recent efforts of the DA community have been  
46 made to counteract this lack in the gridded SSH products, amongst them the dynamical optimal  
47 interpolation (DOI) (Ubelmann et al. 2015), Multiscale Interpolation Ocean Science Topography  
48 (MIOST) (Ubelmann et al. 2021) or BFN-QG (Le Guillou et al. 2023).

49  
50 From another point of view, deep learning frameworks are currently knowing an intense  
51 period of scientific contributions to revisit statistical methods with neural network formulation.  
52 State-of-the-art methods leverage DL to better extract the information from the observations  
53 compared to the classical DA and OI approaches, in which fine-tuning of the DA scheme  
54 parametrizations (background and observation error covariances) demands itself some method-  
55 ological and computational effort, see e.g. Tandeo et al. (2020), all together with the popular  
56 Gaussian simplifications involved in most of DA operational schemes (Asch et al. 2016), though  
57 some approaches exist to alleviate the Gaussian assumptions in nonlinear geophysical problems,  
58 see e.g. Kurosawa and Poterjoy (2023). End-to-end learning architectures are also designed being  
59 backboneed on DA schemes (Boudier et al. 2023; Rozet and Louppe 2023; Fablet et al. 2021), so

60 that they draw from the bayesian formalism to learn all the components of the DA procedure (prior  
61 model, observation operator, numerical solver, etc.) at once. Last, over the last few years, the  
62 successfull applications of generative models to geophysical datasets paves the way to combine  
63 DL together with UQ by efficient sampling strategies of the posterior distribution. Though, to our  
64 knowledge, no end-to-end-combination of DA, DL and UQ has yet been proposed.

65  
66 That is why in this work, we draw from preliminary neural schemes inspired by variational data  
67 assimilation, see e.g. (Fablet et al. 2021), to jointly learn the SPDE parametrization of a surrogate  
68 stochastic prior model of the evolution equation, together with the solver of the minimization  
69 problem. Because the parameters of the SPDE remain initially unknown, they are embedded in the  
70 optimization process using an augmented state formulation, commonly used in data assimilation,  
71 see e.g. (Ruiz et al. 2013). The solver is still based on an iterative residual scheme (Fablet et al.  
72 2021) to update the analysis state. Here, the analysis stands for the expectation of the state given the  
73 observations, together with the SPDE parametrization maximizing their likelihood given the true  
74 states used during the training. The SPDE equation can then be seen as a tangent linear model of the  
75 prior along the data assimilation window, from which we provide uncertainty quantification (UQ)  
76 throughout its precision matrix. Using such a stochastic prior entails the possibility of generating  
77 huge members in the posterior pdf, after conditioning of the prior samples by our neural variational  
78 scheme. Also, if the training dataset is large enough, the method provides an efficient way to  
79 estimate *online* the SPDE parametrization for any new sequence of input observations, without any  
80 additional inference to make. In the end, the key contributions of this work are four-fold:

- 81 • We develop the explicit solver of the considered SPDE prior. It relies on the analytical  
82 expression for the SPDE-based precision matrix of any state trajectory, based on a finite-  
83 difference discretization of the grid covered by the tensors involved in our neural scheme;
- 84 • We exploit this SPDE parametrization as surrogate prior model in the proposed variational  
85 formulation and leverage a trainable gradient-based solver to address jointly the interpolation  
86 of the state trajectory and the estimation of SPDE parameters from irregularly-sampled obser-  
87 vations. The end-to-end training of the solver targets both the expectation of the state given  
88 the observations, together with the SPDE parametrization maximizing its likelihood given the  
89 true states;

- The SPDE prior paves the way to uncertainty quantification through the sampling of the prior pdf and the conditioning by the neural gradient-based solver;
- Two applications of this framework are provided: first, a GP driven by a spatio-temporal SPDE with spatially varying diffusion parameters is compared to the optimal solution; then a Sea Surface Height realistic dataset is used to demonstrate how the proposed framework is also relevant for non-gaussian and non-linear dynamics.

To present these contributions, the paper is structured as follows. In Section 2, we provide a preliminary background on SPDE-based Optimal Interpolation, Data assimilation and Deep Learning for DA, that we aims to bridge together in this work. We also remind how UQ is usually handled in this different fields. In Section 3, we present how we extend neural solvers to embed SPDE prior parametrizations for UQ. Finally, Section 4 provides two applications of this framework in the Gaussian case and for realistic Sea Surface Height datasets.

## 2. Background: Data Assimilation, Uncertainty Quantification and Machine Learning

In this work, we target the reconstruction of a probabilistic spatio-temporal state sequence  $\mathbf{x} = \{\mathbf{x}_k(\mathcal{D})\}, \mathbf{x}_k \in \mathbb{R}^m$  given the partial and potentially noisy observational dataset  $\mathbf{y}(\Omega) = \{\mathbf{y}_k(\Omega_k)\}, \mathbf{y}_k \in \mathbb{R}^{p_k}$ , where  $\Omega = \{\Omega_k\} \subset \mathcal{D}$  denotes the subdomain with observations and index  $k$  refers to time  $t_k$ . To do so, we aim at bridging Data Assimilation (DA), Uncertainty Quantification (UQ) and deep learning-based (DL) methods, see Fig. 1 for a conceptual illustration of these connections, to propose SPDE-based extensions of the so-called 4DvarNet neural variational scheme (Fablet et al. 2021), as a generic interpolation and short-term forecasting tools. We provide here a brief presentation of these three literatures, focusing only on what is useful in our framework, with an additional description of how UQ is usually handled among these communities. For a more exhaustive review, Cheng et al. (2023) provides a detailed presentation of machine learning techniques with data assimilation and uncertainty quantification for dynamical systems.

### *a. Optimal Interpolation (OI) and Data Assimilation (DA)*

(i) *Classic formulations.* As very basic details to ease the link with the other components of this work, we remind that when no dynamical model is available as a prior information, the

117 covariance-based Optimal Interpolation, see e.g. (Chilès and Delfiner 2012), is given by:

$$\mathbf{x}^\star = \mathbf{P}_{\mathbf{xy}} \mathbf{P}_{\mathbf{yy}}^{-1} \mathbf{y} \quad (1)$$

118 where  $\mathbf{P}_{\mathbf{xy}}$  and  $\mathbf{P}_{\mathbf{yy}}$  are covariance matrices coming from the covariance  $\tilde{\mathbf{P}}$  of the observation and  
 119 state vector  $\begin{bmatrix} \mathbf{y} & \mathbf{x} \end{bmatrix}^\top$ :

$$\tilde{\mathbf{P}} = \begin{bmatrix} \mathbf{P}_{\mathbf{xx}} & \mathbf{P}_{\mathbf{xy}} \\ \mathbf{P}_{\mathbf{xy}}^\top & \mathbf{P}_{\mathbf{yy}} \end{bmatrix} \quad (2)$$

120 Broadly speaking, when the prior information is available, typically as high dimensional numerical  
 121 models in geosciences, see e.g. Carrassi et al. (2018), two main categories of DA (Evensen 2009;  
 122 Evensen et al. 2022) exists: variational and sequential methods. They both aims at minimizing  
 123 some energy or functional involving an equation-based dynamical prior and an observation term.  
 124 Drawing from the link established with Gaussian Processes, we can also consider the case of  
 125 noisy observations and ease the link with data assimilation formalism, see Särkkä and Hartikainen  
 126 (2012); Särkkä et al. (2013); Grigorievskiy et al. (2016). The state space model corresponding to  
 127 the GP regression problem writes:

$$\begin{cases} \mathbf{x}_{t+dt} & = \mathbf{M}_{t+dt} \mathbf{x}_t + \boldsymbol{\eta}_t \\ \mathbf{y}_t & = \mathbf{H}_t \mathbf{x}_t + \boldsymbol{\varepsilon}_t \end{cases} \quad (3)$$

128 where  $\boldsymbol{\eta}_t$  is the  $m$ -dimensional noise process and the evolution equation is defined by the feedback  
 129 linear operator matrix  $\mathbf{M}_{t+dt}$ .  $\mathbf{H}_t$  is the observation operator at time  $t$  mapping the state space  
 130 to the observation space and  $\boldsymbol{\varepsilon}_t$  the observational error with covariance matrix  $\mathbf{R}_t$ . Based on  
 131 this time-dependent notations, we also consider global observation operator  $\mathbf{H}$  with dimensions  
 132  $(L \times p_k) \times (L \times m)$  and global observational error covariance matrix  $\mathbf{R}$  with dimensions  $(L \times p_k) \times$   
 133  $(L \times p_k)$  as block diagonal matrices whose each block respectively contains the time-dependent  
 134 observation operator and observational error covariance matrix  $\mathbf{H}_t$  and  $\mathbf{R}_t$ .

135 (ii) *The Stochastic Partial Differential Equation (SPDE) approach in DA.* The Optimal In-  
 136 terpolation implies to factorize dense covariance matrices which is an issue when the size of

137 spatio-temporal datasets is large. Reduced rank approximations, see e.g. (Cressie and Wikle 2015)  
 138 have already been investigated to tackle this specific problem. More recently, the use of sparse  
 139 precision matrices has also been proposed by using tapering strategies (Furrer et al. 2006; Bolin  
 140 and Wallin 2016) or by making use of the link seen by Lindgren et al. (2011) between Stochastic  
 141 Partial Differential Equations (SPDE) and Gaussian Processes. For the latter, if the original link  
 142 was made through the Poisson SPDE equation (Whittle 1953):

$$(\kappa^2 - \Delta)^{\alpha/2} X = \tau Z ; \quad (4)$$

143 where  $\Delta = \sum_{i=1}^d \frac{\partial^2}{\partial s_i^2}$  denotes the Laplacian operator,  $Z$  is a standard Gaussian white noise,  $\kappa = 1/a$ ,  
 144  $a$  denotes the range of the GPX,  $\alpha = \nu + d/2$  and  $\tau$  relates to the marginal variance of  $X$ . It can  
 145 be extended to more complex linear SPDE involving physical processes like advection or diffusion  
 146 (Lindgren et al. 2011; Fuglstad et al. 2015a; Clarotto et al. 2022). The SPDE-based OI formulation  
 147 uses precision matrix formalism, as the inverse of the covariance matrix  $\mathbf{Q} = \tilde{\mathbf{P}}^{-1} = (dxdy)/\tau^2 \cdot \mathbf{B}^T \mathbf{B}$ ,  
 148 see Eq. (2) where  $\mathbf{B}$  is the discretized version of the fractional differential operator  $(\kappa^2 - \Delta)^{\alpha/2}$ ,  
 149 and  $dx, dy$  are the spatial grid step sizes:

$$\mathbf{x}^* = -\mathbf{Q}_{xx}^{-1} \mathbf{Q}_{xy} \mathbf{y} \quad (5)$$

150 By construction,  $\mathbf{Q}$  is sparse which means that we solve a system with sparse Cholesky of complexity  
 151  $O(m^{3/2})$ , while the general Cholesky algorithm is of complexity  $O(m^3)$ . Thus, it opens new avenue  
 152 to cope with massive observational datasets in geosciences while making use of the underlying  
 153 physics of such processes. Let note that the so-called SPDE-based approach can also be used  
 154 as a general spatio-temporal model, even if it is not physically motivated, since it provides a  
 155 flexible way to handle local anisotropies of a large set of geophysical processes. It has known  
 156 numerous applications in the past few years, see e.g. Sigrist et al. (2015); Fuglstad et al. (2015b).  
 157 Though, when considering a spatio-temporal advection-diffusion SPDE, the parameters generally  
 158 vary continuously across space and/or time making their estimation an additional problem to the  
 159 original interpolation task. This estimation generally relies on off-line strategies (Fuglstad et al.  
 160 2015b) embedded in hierarchical models, which can be another computational issue while the set  
 161 of parameters estimated does not automatically transfer to a new dataset.



162 Regarding the transfer of SPDE formulation in the DA formalism, rewriting the covariance-based  
 163 Eq. (2) in terms of precision matrix leads to:

$$\tilde{\mathbf{P}} = \begin{bmatrix} \mathbf{P}_{\mathbf{xx}} & \mathbf{P}_{\mathbf{xx}}\mathbf{H}^T \\ \mathbf{H}\mathbf{P}_{\mathbf{xx}} & \mathbf{H}\mathbf{P}_{\mathbf{xx}}\mathbf{H}^T + \mathbf{R} \end{bmatrix}, \quad \tilde{\mathbf{Q}} = \begin{bmatrix} \mathbf{Q}_{\mathbf{xx}} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} & -\mathbf{H}^T\mathbf{R}^{-1} \\ -\mathbf{R}^{-1}\mathbf{H} & \mathbf{R}^{-1} \end{bmatrix} \quad (6)$$

164 which gives an other version of Eq. (5):

$$\mathbf{x}^* = \left( \mathbf{Q}_{\mathbf{xx}} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} \right)^{-1} \mathbf{H}^T\mathbf{R}^{-1}\mathbf{y} \quad (7)$$

165 whose posterior precision matrix of state  $\mathbf{x}^*$  is  $\mathbf{Q}(\mathbf{x}|\mathbf{y}) = \mathbf{Q}_{\mathbf{xx}} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}$ . This type of formulation  
 166 makes the link between DA framework and SPDE-based formalism maing extensive use of precision  
 167 matrix formulations.

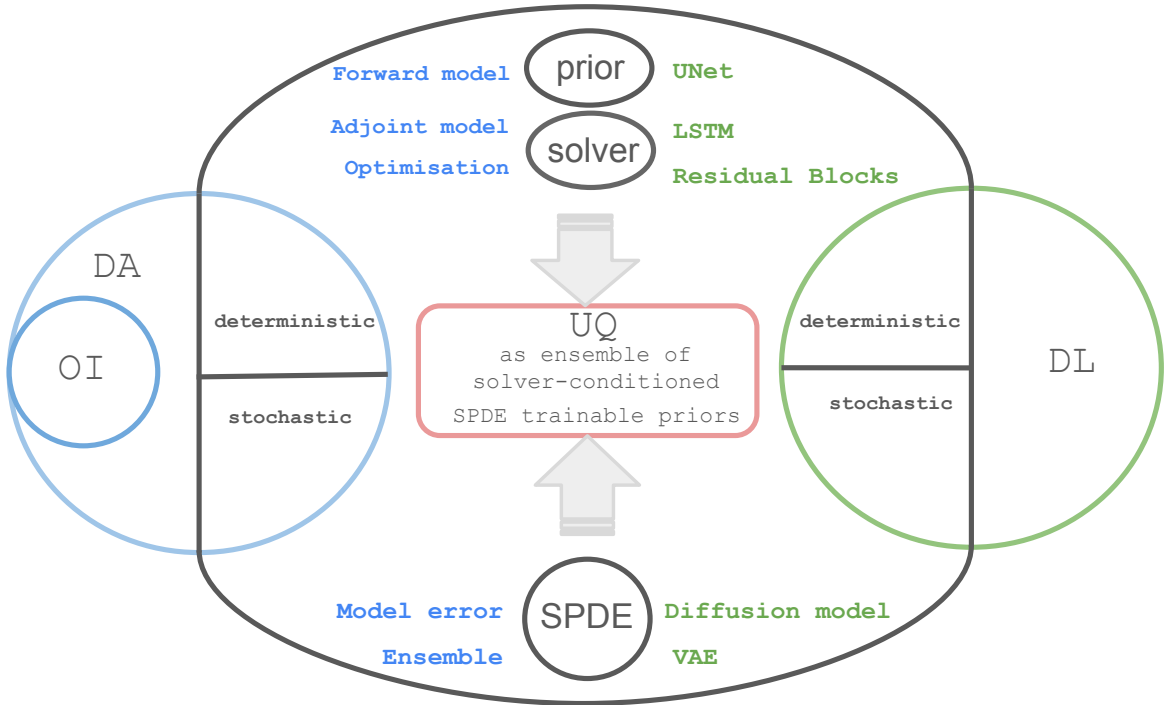


FIG. 1: Combination of SPDE-based formalism, Data Assimilation and Deep Learning to address the challenge of providing UQ in neural variational scheme. We decompose both DA and DL in deterministic and stochastic components to understand how our approach relates to state-of-the-art methods in these fields and how we provide UQ as ensemble of solver-conditioned trainable SPDE priors.

168 *b. Machine and Deep Learning (ML/DL)*

169 From another point of view, deep learning frameworks are currently knowing an intense  
170 period of scientific contribution to revisit statistical methods with neural network formulation.  
171 The latter enables to use automatic differentiation embedded in the gradient-based optimization  
172 as a way of solving traditional inverse problems. Several approaches have been investigated,  
173 among others we can think of using DL as a substitute for one component of the DA procedure:  
174 surrogate dynamical models making extensive use of state-of-the-art neural networks such as  
175 LSTM (Nakamura et al. 2021), UNet (Doury et al. 2023) or Transformers (Tong et al. 2022),  
176 reduced order model to compute DA schemes in latent space or with super-resolution component  
177 (Brajard et al. 2021), modeling of model/observation error covariance (Cheng and Qiu 2022;  
178 Sacco et al. 2022, 2024). The other way rely on end-to-end learning of the entire DA system  
179 instead of using DL techniques to address one aspect of DA algorithm among its three main  
180 blocks: forward model and its error, observation operator and its error, and DA scheme (EnKF,  
181 4DVar, etc.). Such approaches may rely on applying state-of-the-art neural architectures to map  
182 observation data to the hidden state sequence, see e.g. UNet (Li et al. 2022) ,Transformers (Shi  
183 et al. 2022) or LSTM architectures (Martin et al. 2023). In particular, when the problem relates  
184 to space-time interpolation of partial and noisy observation of geophysical fields, DA-inspired  
185 neural schemes have been recently proposed, see e.g. Boudier et al. (2023); Rozet and Louppe  
186 (2023); Fablet et al. (2021). The latter specifically suggests a joint learning of prior models and  
187 solvers as an end-to-end data-driven variational data assimilation scheme. The so-called 4DVarNet  
188 neural scheme is introduced: it involves an implicit iterative gradient-based LSTM solver to  
189 minimize a variational cost, close to what is encountered in 4DVar data assimilation (Carrassi  
190 et al. 2018). In this variational cost, the dynamical prior is no longer equation-based but is stated  
191 as a trainable neural network learnt during the training process. Then, automatic differentiation  
192 is used to compute the gradient of the variational cost during the gradient-based iterations,  
193 instead of requiring the computation of complex and costly adjoint models (Asch et al. 2016).  
194 Drawing from this framework, a neural optimal interpolation scheme has also been proposed  
195 (Beauchamp et al. 2022) to reach OI performance with a linear scaling of the solution on the  
196 number of space-time variables, leading to a significant speed up in the computation of the solution.

197

198 *c. Uncertainty Quantification*

199 (i) *UQ in DA.* Based on its bayesian formalism, Optimal Interpolation provides uncertainty  
200 through a posterior covariance matrix  $\mathbf{P}_{x|y}$ . Because the background  $\mathbf{x}^b$  is generally considered  
201 as stationary, this covariance matrix is mainly driven by the sampling of the observation, see  
202 e.g. (Zhen et al. 2020), which may not be realistic for dynamical systems. Relying on a similar  
203 framework, DA schemes provide the deterministic part of the evolution equation in Eq. 3 as  
204 numerical model outputs, and another probabilistic model has to be given for the distribution of  
205 the noise. In addition, under Gaussian assumption of the model likelihood, closed forms exist for  
206 the mean and standard deviation of the posterior prediction. Ensemble-based sequential schemes  
207 also provide a simple way to compute flow-dependant empirical posterior covariance (Song et al.  
208 2013) that counteracts the sampling issue used in OI or in most of variational schemes. Hybrid  
209 methods in DA (Asch et al. 2016) aims at combining both ensemble-based methods together with  
210 variational schemes to benefit from the assets of each methods.

211 (ii) *UQ for ML approaches.* As stated above, DL for UQ generally relies on the estimation of the  
212 stochastic components of the DA scheme, namely the covariance model and/or observation errors.  
213 More recently, ML schemes aims at substituting the entire DA scheme, either with an explicit  
214 model of the posterior distribution, or by a sampling strategy of the latter. On the first approach,  
215 because the true posterior is both computationally and analytically intractable, a popular strategy is  
216 to estimate an approximate posterior distribution model with trainable parametrization that involves  
217 the minimization of the Kullback-Leibler Divergence between the two distributions, which is the  
218 similar that maximizing the Evidence Lower Bound (ELBO) (Huang et al. 2019) referred as  
219 variational inference (Zhang et al. 2021). More recently, generative models are extensively used  
220 to draw samples in the prior distribution. Among them, GANs (Goodfellow et al. 2014), VAEs  
221 (Kingma and Welling 2022), normalizing flows (Dinh et al. 2017) and diffusion models (Ho et al.  
222 2020) are the most popular. Once a way of sampling the prior is available, the posterior pdf can be  
223 obtained after conditioning through a classic DA scheme or method-related conditional generative  
224 models. This is the case in score-based data assimilation (Rozet and Louppe 2023) stated as  
225 diffusion models where Langevin iterative optimisation scheme can also embed the observation  
226 likelihood (Ho et al. 2020) to give a direct access to the posterior.

### 227 3. Neural variational schemes with SPDE priors

228 We present here how we draw from the neural variational scheme of Fablet et al. (2021) to embed  
 229 an augmented state with SPDE prior parametrization. We explain how to parametrize the SPDE  
 230 with advection-diffusion schemes compliant with many geophysical processes, and how the latter  
 231 leads to the use of sparse prior precision matrix, which is a key aspect in our use of SPDE priors  
 232 for surrogate models in DA. A full overview of the neural solver is also presented together with  
 233 its learning scheme. Last, a presentation of the UQ scheme to generate huge ensembles of the  
 234 posterior pdf is given.

#### 235 *a. Neural MAP solver*

236 Because data assimilation is fundamentally bayesian, most of the methods used to interpolate an  
 237 observational dataset involve the use of a model prior  $\mathbf{x}$  and the computation of the posterior  $E[\mathbf{x}|\mathbf{y}]$   
 238 given the observations. For the latter, the computational time might be expensive, even prohibitive,  
 239 because it implies to solve linear systems with matrices in high dimensions. Recently, alternate  
 240 solutions have been proposed to compute the posterior. Rather than using linear algebra, we can  
 241 use a traditional variational data assimilation scheme (Asch et al. 2016) and the state analysis  
 242  $\mathbf{x}^* = E[\mathbf{x}|\mathbf{y}]$  is obtained by solving the minimization problem:

$$\mathbf{x}^* = \arg \min_{\mathbf{x}} \mathcal{J}(\mathbf{x})$$

243 where the variational cost function  $\mathcal{J}(\mathbf{x}) = \mathcal{J}_{\Phi}(\mathbf{x}, \mathbf{y}, \Omega)$  is generally the sum of an observation term  
 244 and a regularization term involving an operator  $\Phi$  which is typically a dynamical prior:

$$\begin{aligned} \mathcal{J}_{\Phi}(\mathbf{x}, \mathbf{y}, \Omega) &= \mathcal{J}^o(\mathbf{x}, \mathbf{y}, \Omega) + \mathcal{J}_{\Phi}^b(\mathbf{x}) \\ &= \|\mathcal{H}(\mathbf{x}) - \mathbf{y}\|_{\Omega}^2 + \lambda \|\mathbf{x} - \Phi(\mathbf{x})\|^2 \end{aligned} \quad (8)$$

245 with  $\mathcal{H}$  the observation operator and  $\lambda$  is a predefined or learnable scalar weight. This formulation  
 246 of functional  $\mathcal{J}_{\Phi}(\mathbf{x}, \mathbf{y}, \Omega)$  directly relates to weakly constraint 4DVar, see e.g. Carrassi et al. (2018).  
 247 When both prior  $\Phi(\mathbf{x}) = \mathbf{x}^b$  and observations  $\mathbf{y}$  are assumed to be Gaussian with covariance matrices

248  $\mathbf{B}$  and  $\mathbf{R}$ ,  $\mathcal{J}_\Phi(\mathbf{x}, \mathbf{y}, \Omega)$  can be written:

$$\mathcal{J}_\Phi(\mathbf{x}, \mathbf{y}, \Omega) = (\mathbf{H}\mathbf{x} - \mathbf{y})^\top \mathbf{R}^{-1} (\mathbf{H}\mathbf{x} - \mathbf{y}) + \lambda (\mathbf{x} - \mathbf{x}^b)^\top \mathbf{B}^{-1} (\mathbf{x} - \mathbf{x}^b) \quad (9)$$

249 This is well known that equating to zero the gradient of this cost function at a single time  $t_k$   
 250 produces the exact same analysis formulation that the Kalman Filter analysis step or the BLUE  
 251 (simple kriging) equations. For inverse problems with time-related processes, the minimization  
 252 of functional  $\mathcal{J}_\Phi$  usually involves iterative gradient-based algorithms and in particular request to  
 253 consider the adjoint method in classic equation-based variational data assimilation schemes (Asch  
 254 et al. 2016) where operator  $\Phi$  identifies to a deterministic model  $\mathbf{x}_{k+1} = \mathcal{M}(\mathbf{x}_k)$ :

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \alpha \nabla_{\mathbf{x}} \mathcal{J}_\Phi(\mathbf{x}^{(i)}, \mathbf{y}, \Omega)$$

255 Fablet et al. (2020) shows that the so-called 4DVarNet scheme, an end-to-end deep learning  
 256 framework can be built based on the above variational formulation where both prior operator  
 257  $\Phi$  and Maximum a posterior (MAP) solver  $\Gamma$ , i.e. the operator solving for the gradient-based  
 258 minimization of the variational cost, are neural networks. For the latter, following meta-learning  
 259 schemes (Andrychowicz et al. 2016), a residual LSTM-based representation of operator  $\Gamma$  is  
 260 considered where the  $i^{th}$  iterative update of the solver is given by:

$$\begin{cases} \mathbf{g}^{(i+1)} &= LSTM[\alpha \cdot \nabla_{\mathbf{x}} \mathcal{J}_\Phi(\mathbf{x}^{(i)}, \mathbf{y}, \Omega), h(i), c(i)] \\ \mathbf{x}^{(i+1)} &= \mathbf{x}^{(i)} - \mathcal{T}(\mathbf{g}^{(i+1)}) \end{cases} \quad (10)$$

261 with  $\mathbf{g}^{(i+1)}$  is the LSTM output using as input gradient  $\nabla_{\mathbf{x}} \mathcal{J}_\Phi(\mathbf{x}^{(i)}, \mathbf{y}, \Omega)$ , while  $h(i)$  and  $c(i)$   
 262 denotes the internal states of the LSTM (Arras et al. 2019),  $\alpha$  is a normalization scalar and  $\mathcal{T}$  a  
 263 linear or convolutional mapping.

264  
 265 In this formulation, the prior term  $\Phi$  is jointly estimated so that the reconstruction fulfills at most  
 266 the loss function  $\mathcal{L}$  used during the training process. Typically, such a loss function may be stated

267 as:

$$\mathcal{L}(\mathbf{x}, \mathbf{x}^*) = \|\mathbf{x} - \mathbf{x}^*\|^2 + \mathcal{L}_{\text{regul}}, \quad (11)$$

268 the mean squared error between the reconstruction and the ground truth, with additional regular-  
269 ization terms, depending on both the specific application and the targeted outputs of the end-to-end  
270 neural scheme. This implies a bi-level optimization scheme in which we refer to Eqs. (8) and (11)  
271 resp. as the inner variational cost and the outer training loss function of 4DVarNet schemes. When  
272 replacing the outer training loss function by the same inner variational cost, such a scheme would  
273 lead to learn the exact 4DVar scheme, but with additional speeding up the optimization process.  
274 Instead, using this dual optimization enables to escape from the Gaussian and Linear hypotheses  
275 used in the 4DVar formalism so that the training of (11) acts as an additional constraint in the  
276 minimization process of (8).

### 277 *b. SPDE parametrization*

278 Though, the prior  $\Phi$  is not easily interpretable: it acts as an encoding of the state  $\mathbf{x}$  that helps in  
279 the gradient-based minimization process. In this work, we aim at bringing both explainability and  
280 stochasticity in the neural scheme by considering as a surrogate model for prior  $\Phi$  an stochastic  
281 PDE (SPDE) instead of a given neural-based architecture. The continuous process associated with  
282 our state space describes the dynamical evolution of the spatio-temporal prior process  $X_{\theta}(\mathbf{s}, t)$  as  
283 a spatio-temporal SPDE embedding the estimation of its parametrization  $\theta$ , the latter controlling  
284 key physical behaviours such as local anisotropy, correlation range, and marginal variance:

$$\mathcal{F}_{t, \theta(t)}\{X(\mathbf{s}, t)\} = \tau(\mathbf{s}, t)Z(\mathbf{s}, t) \quad (12)$$

285 where operator  $\mathcal{F}_{t, \theta(t)}$  is here considered as linear, then the solution of Eq. (12) is a spatio-temporal  
286 Gaussian Process (Lindgren et al. 2011). Regarding the right-hand side noise of the SPDE, it is  
287 assumed separable, i.e.  $Z(\mathbf{s}, t) = Z_t(t) \otimes Z_s(\mathbf{s})$  with  $Z_t(t)$  a temporal white noise.  $Z_s(\mathbf{s})$  can also  
288 be a spatial white noise or we may consider a colored noise to ensure more regularity on process  
289  $X$  across time.

290

291 Following a state space formalism, the discretization of the stochastic process  $X(\mathbf{s}, t)$  is a multi-  
 292 variate gaussian vector  $\mathbf{x}$  :

$$\mathbf{x} \sim \mathcal{N}(\mathbf{x}^b, \mathbf{Q}_\theta^b) \quad (13)$$

293 where  $\mathbf{x}^b$  is the deterministic mean of the state, typically *the background*, a coarse approximation  
 294 of the field in stationary formulations, or the *forecast* in a dynamical data assimilation scheme.  
 295  $\mathbf{Q}_\theta^b$  is the precision matrix (inverse of the covariance matrix) of the state sequence  $\{\mathbf{x}_0, \dots, \mathbf{x}_{Ldt}\}$ .  
 296 Because of the explicit link between linear stochastic PDEs and Gaussian Processes, we modify  
 297 the variational formulation used in 4DVarNet schemes, see e.g. Beauchamp et al. (2023c), by  
 298 rewriting the matrix formulation of the regularization prior term of Eq. 8 as:

$$\mathbf{x}^\star = \operatorname{argmin}_{\mathbf{x}} \mathcal{J}(\mathbf{x}, \mathbf{y}, \mathcal{Q}) = \operatorname{argmin}_{\mathbf{x}} \|\mathbf{H}\mathbf{x} - \mathbf{y}\|_{\mathcal{Q}}^2 + \lambda [\mathbf{x} - \mathbf{x}^b]^\top \mathbf{Q}_\theta^b [\mathbf{x} - \mathbf{x}^b]$$

299 When  $\mathbf{x}^b = \mathbf{0}$  (for simplification) and when denoting  $\mathbf{L}$  the square root of the precision matrix  $\mathbf{Q}_\theta^b$ ,  
 300 the identification with 4DVarNet schemes is direct for  $\Phi = (1 - \mathbf{L})$ . Indeed,  $\|\mathbf{x}\|_{\mathbf{Q}_\theta^b}^2 = \mathbf{x}^\top \mathbf{L}^\top \mathbf{L} \mathbf{x} =$   
 301  $\|\mathbf{L}\mathbf{x}\|^2 = \|\mathbf{x} - \Phi \cdot \mathbf{x}\|^2$ . In this derived formulation, it is clear that if the SPDE is known, it can  
 302 be embedded in the inner variational cost used by the LSTM iterative solver to optimize the outer  
 303 training loss function. Drawing from the usual neural variational framework, the trainable prior is  
 304 now SPDE-based, and the parameters  $\theta$  are embedded in the following augmented state formalism:

$$\tilde{\mathbf{x}} = \begin{bmatrix} \mathbf{x} & \theta \end{bmatrix}^\top \quad (14)$$

305 The latent parameter  $\theta$  is potentially non stationary in both space and time and its size is directly  
 306 related to the size of the data assimilation window  $L$ .

307 When dealing with geophysical fields, a generic class of non-stationary models generated by  
 308 stochastic PDEs shall introduce some diffusion and/or advection terms. They are respectively  
 309 obtained by introducing a local advection operator  $\mathbf{m}(\mathbf{s}, t) \cdot \nabla$  where  $\mathbf{m}$  is a velocity field and a local

310 diffusion operator  $\nabla \cdot \mathbf{H}(\mathbf{s}, t) \cdot \nabla$  where  $\mathbf{H}$  acts as a two-dimensional diffusion tensor:

$$\frac{\partial X}{\partial t} + \left\{ \kappa^2(\mathbf{s}, t) + \mathbf{m}(\mathbf{s}, t) \cdot \nabla - \nabla \cdot \mathbf{H}(\mathbf{s}, t) \cdot \nabla \right\}^{\alpha/2} X(\mathbf{s}, t) = \tau(\mathbf{s}, t) Z(\mathbf{s}, t) \quad (15)$$

311 This way of handling spatio-temporal non-stationarities in SPDE models has been first men-  
 312 tioned in the original paper of Lindgren et al. (2011), but because of the challenge of estimating  
 313 the full set of space-time parameters, no works have been published to our knowledge pushing  
 314 this framework into its completeness. Though, some parametrization of purely spatial diffusion  
 315 process (Fuglstad et al. 2015c) or stationary advection-dominated SPDE (Clarotto et al. 2022) has  
 316 been successfully applied.

317 Let stress that the advection-diffusion scheme is a good candidate for many geophysical datasets:  
 318 this is the case in quasi-geostrophic approximation of Sea Surface Temperature (SST), see e.g.  
 319 (Ubelmann et al. 2014), but also in the dispersion of atmospheric pollutants (Menut et al. 2021) for  
 320 instance. Then, this framework provides a generic and convenient way to bring more explainability  
 321 in terms of space-time covariances of dynamical processes.

322  
 323 Such a model implies to estimate new parameters  $\kappa(\mathbf{s}, t)$ ,  $\mathbf{H}_{2 \times 2}(\mathbf{s}, t)$  and  $\mathbf{m}_{2 \times 1}(\mathbf{s}, t) = \begin{bmatrix} \mathbf{m}^1 & \mathbf{m}^2 \end{bmatrix}^T$ ,  
 324 all varying across space and time along the data assimilation window. In addition,  $\kappa$  needs to  
 325 be continuous while  $\mathbf{m}$  and  $\mathbf{H}$  additionally requires to be continuously differentiable. Regarding  
 326 the diffusion tensor, we draw from the spatial statistics literature, see e.g. (Fuglstad et al. 2015a),  
 327 to introduce the scalars  $\gamma(\mathbf{s}, t)$ ,  $\beta(\mathbf{s}, t)$ ,  $\mathbf{v}_1(\mathbf{s}, t)$  and  $\mathbf{v}_2(\mathbf{s}, t)$  as a generic decomposition of  $\mathbf{H}(\mathbf{s}, t)$   
 328 through the equation:

$$\mathbf{H}(\mathbf{s}, t) = \begin{bmatrix} \mathbf{H}^{1,1} & \mathbf{H}^{1,2} \\ \mathbf{H}^{1,2} & \mathbf{H}^{2,2} \end{bmatrix}(\mathbf{s}, t) = \gamma(\mathbf{s}, t) \mathbf{I}_2 + \beta(\mathbf{s}, t) \mathbf{v} \mathbf{v}^T$$

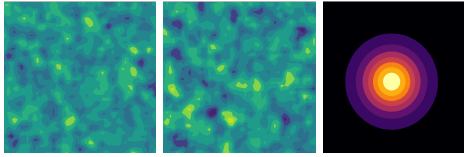
329 with  $\mathbf{v}^T = \begin{bmatrix} \mathbf{v}_1(\mathbf{s}) & \mathbf{v}_2(\mathbf{s}) \end{bmatrix}$ , which models the diffusion tensor as the sum of an isotropic and  
 330 anisotropic effects, the latter being described by its amplitude and magnitude. This is a valid  
 331 decomposition for any symmetric positive-definite  $2 \times 2$  matrix. This leads to the SPDE hyper-  
 332 parametrization  $\theta$  of size  $m \times L \times 8$  parameters ( $\mathbf{H}^{1,2} = \mathbf{H}^{2,1}$ ): it grows linearly with the potentially



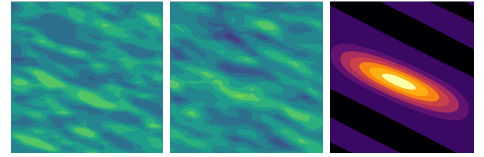
333 high dimensional state space. In the end, the SPDE hyperparametrization states as:

$$\theta = \left[ \kappa \quad \mathbf{m} \quad \mathbf{H} \quad \tau \right]^T$$

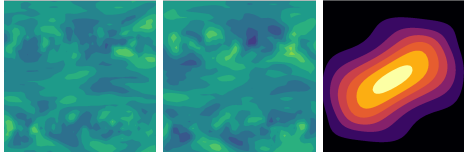
334 Fig.2 provide some specific examples to show how the modifications in the fractional differential  
 335 operator leads to more complex spatio-temporal anisotropies. In this four SPDE parametrizations,  
 336  $\kappa = 0.33$ ,  $\tau = 1$  and  $\alpha = 4$ .



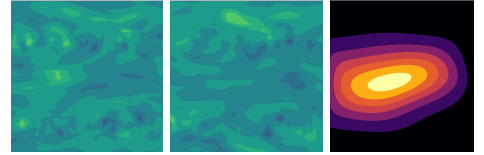
(a) Isotropic model:  $\frac{\partial \mathbf{x}}{\partial t} + (\kappa^2(\mathbf{s}, t) - \Delta)^{\alpha/2} \mathbf{x}(\mathbf{s}, t) = \tau \mathbf{z}(\mathbf{s}, t)$



(b) Global anisotropy:  $\frac{\partial \mathbf{x}}{\partial t} + \{\kappa^2(\mathbf{s}, t) - \nabla \cdot \mathbf{H} \nabla\}^{\alpha/2} \mathbf{x}(\mathbf{s}, t) = \tau \mathbf{z}(\mathbf{s}, t)$



(c) Local anisotropy with diffusion:  $\frac{\partial \mathbf{x}}{\partial t} + \{\kappa^2(\mathbf{s}, t) - \nabla \cdot \mathbf{H}(\mathbf{s}) \nabla\}^{\alpha/2} \mathbf{x}(\mathbf{s}, t) = \tau \mathbf{z}(\mathbf{s}, t)$



(d) Local anisotropy + Global advection:  $\frac{\partial \mathbf{x}}{\partial t} + \{\kappa^2(\mathbf{s}, t) + \mathbf{m} \cdot \nabla - \nabla \cdot \mathbf{H}(\mathbf{s}) \nabla\}^{\alpha/2} \mathbf{x}(\mathbf{s}, t) = \tau \mathbf{z}(\mathbf{s}, t)$

FIG. 2: For (a), (b), (c) and (d), one realization of the corresponding SPDE-driven GP at time  $t = 10$  (left panel),  $t = 20$  (middle panel) and covariance (right panel) with central point of domain  $\mathcal{D} = [0, 1] \times [0, 1]$

337 The space-time SPDE is discretized based on a numerical implicit Euler scheme:

$$\frac{\mathbf{x}_{t+dt} - \mathbf{x}_t}{dt} + \mathbf{B}_{t+dt} \mathbf{x}_{t+dt} = \frac{\tau}{\sqrt{dt}} \mathbf{z}_{t+dt}$$

338 where  $\mathbf{x}_t$  and  $\mathbf{B}_t$  respectively denote the state space and the finite difference discretization of the  
 339 fractional differential operator  $\frac{\partial}{\partial t} + \{\kappa^2(\mathbf{s}, t) + \mathbf{m}(\mathbf{s}, t) \cdot \nabla - \nabla \cdot \mathbf{H}(\mathbf{s}, t) \nabla\}^{\alpha/2}$  at time  $t = 0, \dots, L$ . The  
 340 noise  $\mathbf{z}_{t+dt}$  is white in space and  $\mathbf{M}_{t+dt} = (\mathbf{I} + dt \mathbf{B}_{t+dt})^{-1}$  denotes the matrix operator that emulates  
 341 the dynamical evolution of state  $\mathbf{x}$  from time  $t$  to  $t + dt$ . In this case,  $\mathbf{T}_{t+dt} = \tau \sqrt{dt} \mathbf{M}_{t+dt}$  corresponds  
 342 to the dynamical linear model regularized by the product of the noise variance with the square root  
 343 of the SPDE time step. In case of a more general right-hand term with non-uniform regularization

344 variance  $\{\tau_t, t > 0\}$ , and  $\{\mathbf{z}_t, t > 0\}$  are independent realizations of a colored noise driven by a  
 345 spatial isotropic SPDE:

$$(\kappa_s^2 - \Delta)^{\alpha_s/2} Z_s(\mathbf{s}) = W(\mathbf{s}) \quad (16)$$

346 with  $W(\mathbf{s})$  a white noise with unit variance, the spatial FDM together with time Euler discretization  
 347 leads to:

$$\mathbf{x}_{t+dt} = \mathbf{M}_{t+dt} \mathbf{x}_t + \widetilde{\mathbf{T}}_{t+dt} \mathbf{z}_{t+dt} \quad (17)$$

348 where  $\widetilde{\mathbf{T}}_{t+dt} = \sqrt{dt} \mathbf{M}_{t+dt} \tau_t \mathbf{L}_s$  and  $\mathbf{L}_s$  stands for the Cholesky decomposition of the discretized  
 349 spatial precision matrix  $\mathbf{Q}_s$  introduced of the stochastic process  $Z_s$  introduced by Eq. (16). In  
 350 the case of advection-dominated SPDEs, we involve state-of-the-art upwind schemes (UFDM)  
 351 for stabilization of the numerical system, by letting the advective transport term, which is the  
 352 dominating term, collect its information in the flow direction, i.e., upstream or upwind of the point  
 353 in question. All the calculation details are given in Appendix 5.

354  
 355 In a compact formulation, using centered finite differences on the diffusion term, and by denoting  
 356  $\mathbf{a}_{i,j}^{1,t,+} = \max(\mathbf{m}_{i,j}^{1,t}, 0)$ ,  $\mathbf{a}_{i,j}^{1,t,-} = \min(\mathbf{m}_{i,j}^{1,t}, 0)$ ,  $\mathbf{a}_{i,j}^{2,t,+} = \max(\mathbf{m}_{i,j}^{2,t}, 0)$ ,  $\mathbf{a}_{i,j}^{2,t,-} = \min(\mathbf{m}_{i,j}^{2,t}, 0)$ , the resulting  
 357 UFDM scheme is:

$$\begin{aligned} \mathbf{x}_{i,j}^{t+1} = & \mathbf{x}_{i,j}^t + dt \left[ \kappa_{i,j}^t \mathbf{x}_{i,j}^t + \left( \mathbf{a}_{i,j}^{1,t,+} \mathbf{m}_{i,j}^{1,t,-} + \mathbf{a}_{i,j}^{1,t,-} \mathbf{m}_{i,j}^{1,t,+} \right) + \left( \mathbf{a}_{i,j}^{2,t,+} \mathbf{m}_{i,j}^{2,t,-} + \mathbf{a}_{i,j}^{2,t,-} \mathbf{m}_{i,j}^{2,t,+} \right) \right. \\ & + \mathbf{H}_{i,j}^{1,1,t} \frac{\mathbf{x}_{i+1,j}^t - 2\mathbf{x}_{i,j}^t + \mathbf{x}_{i-1,j}^t}{dx^2} + \mathbf{H}_{i,j}^{2,2,t} \frac{\mathbf{x}_{i,j+1}^t - 2\mathbf{x}_{i,j}^t + \mathbf{x}_{i,j-1}^t}{dy^2} \\ & \left. + \mathbf{H}_{i,j}^{1,2,t} \frac{\mathbf{x}_{i+1,j+1}^t - \mathbf{x}_{i+1,j-1}^t - \mathbf{x}_{i-1,j+1}^t + \mathbf{x}_{i-1,j-1}^t}{2dxdy} + \tau_{i,j}^t \mathbf{z}_{i,j}^{t+1} \right] \end{aligned}$$

358 Starting from this numerical scheme, the modified 4DVarNet scheme requires the precision  
 359 matrix  $\mathbf{Q}_\theta^b$  of the state sequence  $\{\mathbf{x}_0, \dots, \mathbf{x}_{Ldt}\}$ . Here,  $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{P}_0)$  denotes the initial state and  
 360  $\mathbf{Q}_0 = \mathbf{P}_0^{-1}$  is always taken as the precision matrix obtained after a given stabilization run, i.e.  
 361 the evolution of the dynamical system over  $N$  timesteps using as stationary parameters the initial

362 parametrization  $\theta_0$  of the SPDE at time  $t = 0$ , then we can rewrite :

$$\{\mathbf{x}_0, \dots, \mathbf{x}_{Ldt}\} = \mathbf{M}_G \begin{bmatrix} \mathbf{x}_0 \\ \mathbf{z} \end{bmatrix}$$

363 with  $\mathbf{z} = [\mathbf{z}_1, \dots, \mathbf{z}_t]^\top$  and

$$\mathbf{M}_G = \begin{bmatrix} \mathbf{I} & 0 & 0 & 0 & 0 & \dots & 0 \\ \mathbf{M}_1 & \mathbf{T}_1 & 0 & 0 & 0 & \dots & 0 \\ \mathbf{M}_2\mathbf{M}_1 & \mathbf{M}_2\mathbf{T}_1 & \mathbf{T}_2 & 0 & 0 & \dots & 0 \\ \mathbf{M}_3\mathbf{M}_2\mathbf{M}_1 & \mathbf{M}_3\mathbf{M}_2\mathbf{T}_1 & \mathbf{M}_3\mathbf{T}_2 & \mathbf{T}_3 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \mathbf{T}_L \end{bmatrix}$$

364 With the additional notation  $\mathbf{S}_k = \mathbf{T}_k \mathbf{T}_k^\top$ , see Eq. (17), the precision matrix  $\mathbf{Q}^b$  writes, see Appendix  
365 c for all the details:

$$\mathbf{Q}_\theta^b = \frac{1}{dt} \begin{bmatrix} \mathbf{P}_0^{-1} + \tilde{\mathbf{Q}}_{s,1} & -\tilde{\mathbf{Q}}_{s,1}\mathbf{M}_1^{-1} & 0 & 0 & 0 & \dots & 0 \\ -(\mathbf{M}_1^\top)^{-1}\tilde{\mathbf{Q}}_{s,1} & \mathbf{M}_1^\top\tilde{\mathbf{Q}}_{s,1}\mathbf{M}_1 + \tilde{\mathbf{Q}}_{s,2} & -\tilde{\mathbf{Q}}_{s,2}\mathbf{M}_2^{-1} & 0 & 0 & \dots & 0 \\ 0 & -(\mathbf{M}_2^\top)^{-1}\tilde{\mathbf{Q}}_{s,2} & \mathbf{M}_2^\top\tilde{\mathbf{Q}}_{s,2}\mathbf{M}_2 + \tilde{\mathbf{Q}}_{s,3} & -\tilde{\mathbf{Q}}_{s,3}\mathbf{M}_3^{-1} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -(\mathbf{M}_{L-1}^\top)^{-1}\tilde{\mathbf{Q}}_{s,L-1} & \mathbf{M}_L^\top\tilde{\mathbf{Q}}_{s,L-1}\mathbf{M}_{L-1} + \tilde{\mathbf{Q}}_{s,L} & -\tilde{\mathbf{Q}}_{s,L}\mathbf{M}_L^{-1} \\ 0 & \ddots & \ddots & \ddots & 0 & -(\mathbf{M}_L^\top)^{-1}\tilde{\mathbf{Q}}_{s,L} & \mathbf{M}_L^\top\tilde{\mathbf{Q}}_{s,L}\mathbf{M}_L \end{bmatrix} \quad (18)$$

366 where  $\tilde{\mathbf{Q}}_{s,t}$  is the precision matrix of the colored noise weighted by the non-uniform regularization  
367 variance  $\tau_t$ . As clearly visible, the sparsity of  $\mathbf{Q}_\theta^b$  is high, which is key in traditional SPDE-based  
368 GP inference, but also in our approach, see Section c.

369 *c. Neural solver with augmented state*

370 Overall, let denote by  $\Psi_{\theta,\Gamma}(\tilde{\mathbf{x}}^{(0)}, \mathbf{y}, \Omega)$  the output of the end-to-end learning scheme given the  
371 SPDE-based dynamical model with parameters  $\theta$  and the neural residual architecture for the solver

372  $\Gamma$ , see Fig. 3 and Algorithm 1, the initialization  $\tilde{\mathbf{x}}^{(0)}$  of augmented state  $\tilde{\mathbf{x}}$  and the observations  $\mathbf{y}$   
 373 on domain  $\Omega$ .

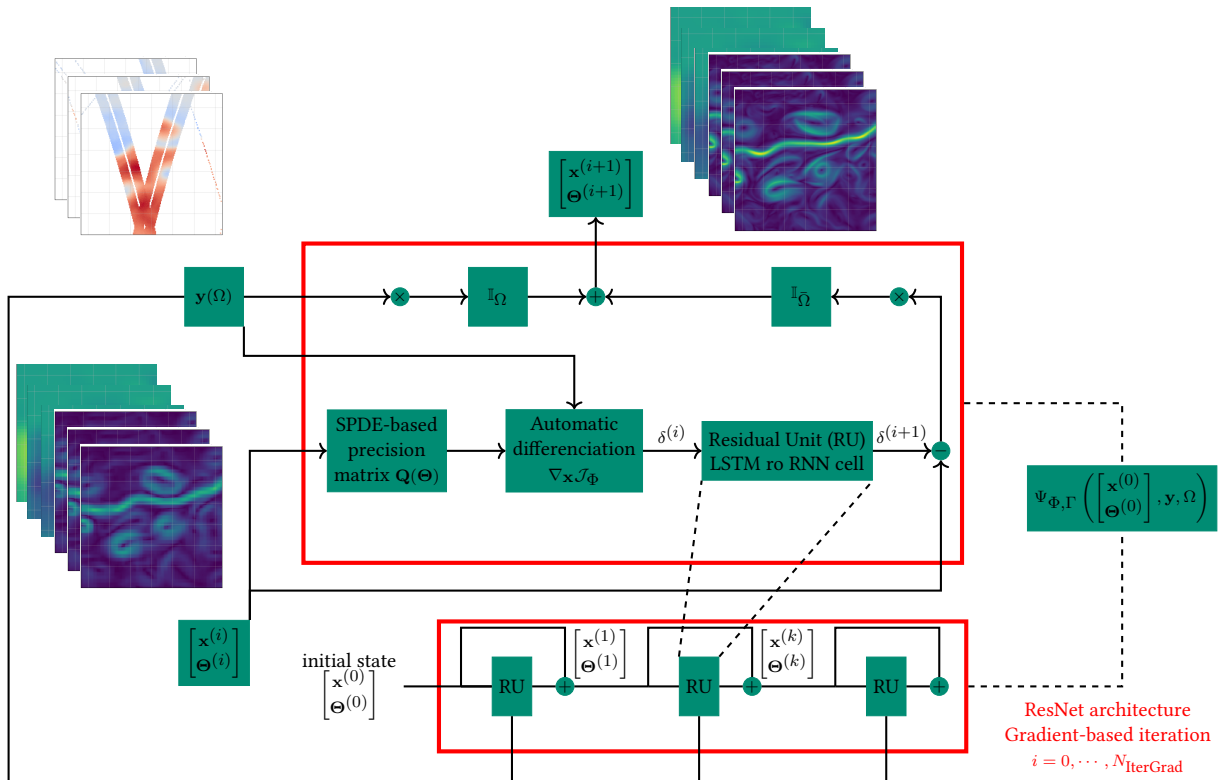


FIG. 3: Sketch of the gradient-based algorithm.  $\mathbf{I}_{\Omega}$  acts as a masking operator for any spatio-temporal location not in  $\Omega$ .

---

**Algorithm 1:** Variational scheme with SPDE-based GP prior and implicit neural solver
 

---

**Data:**
 $\mathbf{x} \in \mathbb{R}^{T \times m} = \{\mathbf{x}_k\}, k = 1, \dots, T$ 
 $\mathbf{y}_\Omega = \{\mathbf{y}_{k, \Omega_k}\}, k = 1, \dots, T$ : observations on domains  $\Omega_k \subset \mathcal{D}$ 
 $N_I$ : number of iterations

 $\eta$ : gradient step

**Init:**
 $\tilde{\mathbf{x}}^{\star, (0)}$ 
**List of procedures:**

 Train\_ $\Psi_{\theta, \Gamma}$ : end-to-end learning procedure with:

 $\theta$ : parameter of the SPDE-based prior operator;

*GradLSTM*: residual NN-based representation of  $\nabla_{\mathbf{x}} \mathcal{J}(\mathbf{x})$ 
 $\Gamma$ : iterative gradient-based update operator:

 $i = 0$ 
**while**  $i < N_I$  **do**

$$\mathbf{Q}_{\theta^{\star, (i)}}^b = \mathbf{P}_G^{-1}(\theta^{\star, (i)}) = \mathbf{M}_G^{-1}(\theta^{\star, (i)})^T \begin{bmatrix} \mathbf{P}_0^{-1} & 0 & \dots & 0 \\ 0 & \mathbf{I} & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{I} \end{bmatrix} \mathbf{M}_G^{-1}(\theta^{\star, (i)})$$

 $\tilde{\mathbf{x}}^{(i+1)} \leftarrow \tilde{\mathbf{x}}^{(i)} - \eta \times \text{GradLSTM}(\tilde{\mathbf{x}}^{(i)})$ 
 $N_I \nearrow; \eta \searrow; i \leftarrow i+1$ 
**end**
**for**  $i \in 0, \dots, n_{epochs}$  **do**
 $\omega_{\Psi_{\theta, \Gamma}}^{(i+1)} \leftarrow \omega_{\Psi_{\theta, \Gamma}}^{(i)} - lr \times \nabla \mathcal{L}(\mathbf{x}, \mathbf{x}^{\star, (i)}, \theta^{(i)})$ 
**end**
**Result:**  $\tilde{\mathbf{x}}^{\star} \leftarrow \Psi_{\theta, \Gamma}(\tilde{\mathbf{x}}^{(0)}, \mathbf{y}, \Omega)$ 


---

374

 375 Then, the joint learning for the weights  $\omega_{\theta, \Gamma}$  of the neural scheme given the SPDE formulation  
 376 that is chosen (isotropic or not, non-stationary or not, etc.) and the architecture of operator  $\Gamma$  is

377 stated as the minimization of the mixed loss function  $\mathcal{L}(\mathbf{x}, \mathbf{x}^*, \boldsymbol{\theta}^*)$ , fully explained in Section d:

$$\omega_{\Psi_{\boldsymbol{\theta}, \Gamma}}^* = \arg \min_{\boldsymbol{\theta}, \Gamma} \left[ \mathcal{L}(\mathbf{x}, \mathbf{x}^*, \boldsymbol{\theta}^*) \right] \text{ s.t. } \tilde{\mathbf{x}}^* = \Psi_{\boldsymbol{\theta}, \Gamma}(\tilde{\mathbf{x}}^{(0)}, \mathbf{y}, \Omega) \quad (19)$$

378 (i) *Initialization of the augmented state.* The initial state  $\tilde{\mathbf{x}}^{(0)}$  should be filled with 0 in the non-  
 379 observed part of domain  $\mathcal{D}$ . Because there is obviously no observation of the SPDE parameters,  
 380 we might guide the training process at the first iteration of the solver. This proved to be particularly  
 381 helpful when dealing with realistic geophysical datasets for which the advection diffusion scheme  
 382 is meaningful. In that case, see for instance Application 2 on real SSH datasets in Section 4, we  
 383 used first and second-order derivatives of the initial state  $\mathbf{x}^{(0)}$  as initial parametrizations for the  
 384 advection and diffusion process:

$$\mathbf{m}^{(0)} = \begin{pmatrix} \frac{\partial \mathbf{x}^{(0)}}{\partial x} \\ \frac{\partial \mathbf{x}^{(0)}}{\partial y} \end{pmatrix}, \text{ and } \mathbf{v}^{(0)} = \begin{pmatrix} \frac{\partial^2 \mathbf{x}^{(0)}}{\partial x^2} \\ \frac{\partial^2 \mathbf{x}^{(0)}}{\partial y^2} \end{pmatrix}$$

385  $\boldsymbol{\kappa}^{(0)}$  and  $\boldsymbol{\tau}^{(0)}$  are both using absolute values of the normalized gradient norms, while  $\beta^{(0)}$  and  $\gamma^{(0)}$   
 386 are resp. set to 1 and 0. Because  $\kappa$ ,  $\tau$  and  $\gamma$  are both strictly positive, see Section b, we use ReLu  
 387 activation function on these three parameters to ensure their consistency.

388 (ii) *Computational aspects.* The sparse formulation of the precision matrix  $\mathbf{Q}_{\boldsymbol{\theta}}^b$  is key in the  
 389 memory-saving component of the algorithm because the latter relies on a set of  $N_I$  gradient-based  
 390 iterations, meaning that for a single interpolation task, the precision matrix  $\mathbf{Q}_{\boldsymbol{\theta}}^b$  is stored  $N_I$  times  
 391 along the computational graph with updated values of the SPDE parameters  $\boldsymbol{\theta}^{(i)}$ .

#### 392 d. Learning scheme

393 (i) *Training loss.* The joint problem of estimating the best reconstruction and inferring realistic  
 394 SPDE parametrizations is difficult because according to the size and nature of the dataset, the  
 395 spatio-temporal interpolation may not always benefit from knowing the exact set of true SPDE  
 396 parameters. Indeed, if the degree of sparsity of the observation dataset is low, the reconstruction  
 397 may be good despite a poor estimation of the covariance matrix. In the other way, if the degree  
 398 of sparsity is high, much more difficult will be the estimation of the SPDE underlying parameters.

399 In this section, we benefit from the supervised configuration of neural variational scheme to train  
400 this joint problem. In this Section, we show how to embed the global precision matrix defined by  
401 Eq. 18 in our neural scheme and define which training loss is the more appropriate to handle the  
402 bi-level optimization scheme (of both the inner variational cost and the training loss).

403 For the training process, we may consider different loss functions:

- 404 •  $\mathcal{L}_1(\mathbf{x}, \mathbf{x}^*) = \|\mathbf{x} - \mathbf{x}^*\|^2$  is the L2-norm of the difference between state  $\mathbf{x}$  and reconstruction  $\mathbf{x}^*$
- 405 •  $\mathcal{L}_2(\mathbf{x}, \theta^*) = -|\mathbf{Q}_{\theta^*}^b| + \mathbf{x}^T \mathbf{Q}_{\theta^*}^b \mathbf{x}$  is the negative log-likelihood of the true states given the esti-  
406 mated precision matrix  $\mathbf{Q}_{\theta^*}^b$ , thus ensuring consistency between the actual ground truth and  
407 the SPDE parameters.

408 Using  $\mathcal{L}_1$  will lead to satisfactory reconstructions without any constraints on the SPDE param-  
409 eters. The single use of  $\mathcal{L}_2$  should lead to satisfactory results if the analytical solver, i.e. the  
410 inversion of the linear system or the gradient-based minimization of the variational cost, were used.  
411 But because the solver is trained, it also needs to be constrained by an appropriate loss function  
412 for the reconstruction, meaning  $\mathcal{L}_1$ . The best solution is then to create a mixed loss function,  
413 combination of  $\mathcal{L}_2$  to estimate at best the SPDE parameters and optimize the prior model, and  $\mathcal{L}_1$   
414 to satisfy the reconstruction criteria and optimize the solver.

415 The log-determinant of the precision matrix  $\log|\mathbf{Q}_{\theta^*}^b|$  is usually difficult to handle when comput-  
416 ing  $\mathcal{L}_2$ . Hopefully, based on the particular structure and the notations already introduced for the  
417 spatio-temporal precision matrix  $\mathbf{Q}_{\theta}^b$  in Eq.(B2), it writes, see e.g. Clarotto et al. (2022):

$$\begin{aligned}
\log|\mathbf{Q}_{\theta^*}^b| &= \log|\mathbf{P}_0^{-1}| + \log|\mathbf{S}_1^{-1}| + \dots + \log|\mathbf{S}_L^{-1}| \\
&= \log|\mathbf{P}_0^{-1}| + \sum_{i=1}^L \log\left(|\mathbf{L}_i \mathbf{L}_i^T|\right) \\
&= \log|\mathbf{P}_0^{-1}| + 2 \sum_{i=1}^L \sum_{j=1}^m \log \mathbf{L}_i(j, j)
\end{aligned} \tag{20}$$

418 where  $\mathbf{L}_i$  denotes here the Cholesky decomposition of  $\mathbf{S}_k^{-1}$ . In case of unsupervised learning, the  
419 same strategy may apply but  $\mathcal{L}_2$  will be the likelihood of the observations given the estimated SPDE  
420 parameters since in this case, the true states would not be available during the training process.

421 (ii) *Two-step learning schemes.* In the initial version of 4DVarNet schemes, the output  $\mathbf{x}^\star$  provided  
 422 by the neural formulation is deterministic and can be seen as the posterior mean of the state given  
 423 the observations. Within this SPDE-based parametrization of the prior  $\mathbf{x}^b$ , we also aim at providing  
 424 the distribution of the prior as a GP process. Though, for realistic geophysical fields, even the prior  
 425 cannot be considered as a zero-mean Gaussian field and some non-linearities have to be accounted  
 426 for in its deterministic mean  $\mathbf{x}^b$ . To solve for this specific case, we involve a 2-step learning process  
 427 in which the deterministic mean  $\mathbf{x}^b$  is first estimated by a 4DVarNet scheme applied on coarser  
 428 resolution than the actual observations. Second, the modified 4DVarNet scheme is involved to  
 429 estimate jointly the posterior mean  $\mathbf{x}^\star$  together with the SPDE parametrization  $\theta^\star$  of the prior  
 430  $\mathbf{x} \sim \mathcal{N}(\mathbf{x}^b, \mathbf{Q}_{\theta^\star}^b)$ . This two-steps procedure also enables to simplify the SPDE training scheme: the  
 431 first guess is estimated based on the entire set of observations along the state sequence, while the  
 432 SPDE parametrization is estimated only on a reduced window of length 5, centered on the targeted  
 433 time of interest. This considerably reduces the size of precision matrix  $\mathbf{Q}_\theta^b$ , making the algorithm  
 434 scalable for any application. Fig. 4 shows a schematic overview of this two-steps learning scheme  
 435 with illustrations coming from realistic Sea Surface Height datasets provided in Application 2,  
 436 Section 4. In the end, this will lead to a stochastic version of the neural variational scheme based  
 437 on the prior GP distribution in which we can sample ensemble members.



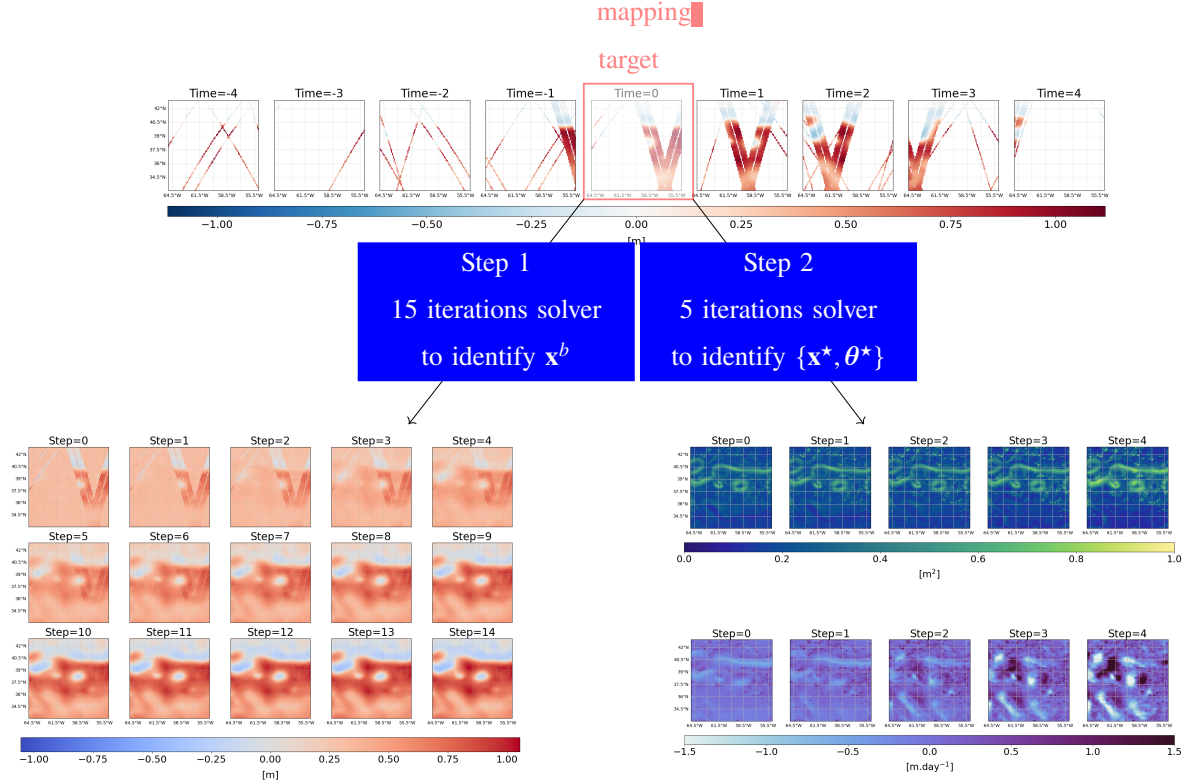


FIG. 4: Two-step adaptation of the 4DVarNet scheme: on the top panel, an example of 4 nadirs+SWOT observations along the 9-day data assimilation window, see Application 2 in Section 4. The target reconstruction day is at the center of the DAW. On the bottom-left panel is displayed the 15 iterations of the first 4DVarNet solver to retrieve the prior mean  $\mathbf{x}^b$ : at the beginning, the non-observed parts of the domain are filled with the global mean of the training dataset (0 when normalized). On the bottom-right panel is displayed the 5 iterations of the modified 4DVarNet solver with augmented state formalism to retrieve both  $\mathbf{x}^*$  and  $\theta^*$ : at the beginning, the parameters in  $\theta$  are initialized with partial derivatives of  $\mathbf{x}^b$ , see Section 3.c

438 (iii) *Complementary PyTorch developments.* Regarding the implementation of our model, we  
 439 use Pytorch (Paszke et al. 2017) whose sparse linear algebra is not providing yet a sparse Cholesky  
 440 algorithm and a sparse solver of linear systems, which is critical, especially when computing the  
 441 likelihood  $\mathcal{L}_2$  in the training loss function. As a consequence, despite the theoretical tools have  
 442 been fully detailed in the previous sections, we had to implement new functionalities based on  
 443 scipy sparse linear algebra (Virtanen et al. 2020) to store the precision matrices in an efficient  
 444 way and compute the inner variational cost, see again Eq.8. In particular, we provide a PyTorch  
 445 extension of sparse cholesky matrices, see Appendix c, based on the CHOLMOD supernodal  
 446 Cholesky factorization (Chen et al. 2008) and draw from Seeger et al. (2019) to provide the

447 backward pass of the sparse Cholesky decomposition.

448

449 *e. UQ scheme*

450 Because the PDE is stochastic, it provides an easy way to generate a set of  $N$  gaussian prior  
451 simulations  $\mathbf{x}_i, i = 1, \dots, N$ :

$$\mathbf{x}_i = \mathbf{x}^b + \mathbf{L}_\theta^b \mathbf{z}_i$$

452 where  $\mathbf{L}_\theta^b$  stands for the Cholesky decomposition of  $\mathbf{Q}_\theta^b$  and  $\mathbf{z}_i$  is a white noise. From a geostatistical  
453 point of view, this can be seen as SPDE-based spatio-temporal non conditional simulations of state  
454  $\mathbf{x}$ , meaning that we produce surrogate simulations sharing the same physical properties than the true  
455 states. These simulations are then conditioned by the neural solver given the observations available  
456  $\mathbf{y}$ , see Fig. 5. To do so, we draw from traditional geostatistics to realize SPDE-based spatio-  
457 temporal non conditional simulations with a kriging-based conditioning (Wackernagel 2003).  
458 Except that we replace the kriging algorithm by our neural approach, which has to be seen as a  
459 generic interpolation tool here:

$$\mathbf{x}_i^\star(\mathbf{s}, t) = \mathbf{x}^\star(\mathbf{s}, t) + (\mathbf{x}_i(\mathbf{s}, t) - \widehat{\mathbf{x}}_i(\mathbf{s}, t)) \quad (21)$$

460 where  $\mathbf{x}^\star$  denotes the neural-based interpolation,  $\mathbf{x}_i$  is one SPDE non-conditional simulation of the  
461 process  $\mathbf{x}$  based on the parameters  $\theta^\star$  and  $\widehat{\mathbf{x}}_i$  is the neural reconstruction of this non-conditional  
462 simulation, using as pseudo-observations a subsampling of  $\mathbf{x}_i$  based on the actual data locations.  
463 Because  $E[\mathbf{x}_i - \widehat{\mathbf{x}}_i] = 0$ , the resulting simulation is well conditioned by the observations at data  
464 locations.

465

466 Running an ensemble of  $N$  conditional simulations gives an approximation of the probability  
467 distribution function  $p_{\mathbf{x}|\mathbf{y}}$  of state  $\mathbf{x}^\star = \mathbf{x}|\mathbf{y} = \{\mathbf{x}_0|\mathbf{y}, \dots, \mathbf{x}_L|\mathbf{y}\}$ . The ensemble mean  $\overline{\mathbf{x}_i^\star}$  will be  $\mathbf{x}^\star$  in

468 the limits of  $N \rightarrow +\infty$ :

$$\frac{1}{N} \sum_i \mathbf{x}_i^*(\mathbf{s}, t) \xrightarrow{N \rightarrow +\infty} \mathbf{x}^*(\mathbf{s}, t)$$

469 Such an approach has already been successfully tested in Beauchamp et al. (2023a) when using  
 470 analog operator strategy (Tandeo et al. 2015) to draw non-conditional simulation in the prior  
 471 distribution.

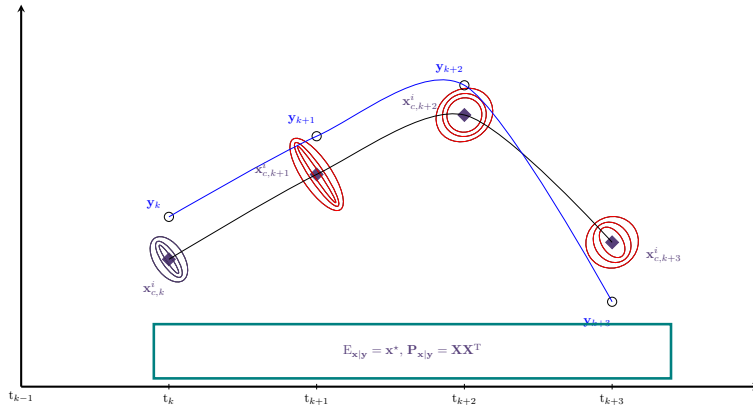


FIG. 5: Ensemble-based neural variational scheme with SPDE-based GP prior: the assimilation is not sequential. The inversion scheme embeds the global precision matrix of the state sequence  $\{\mathbf{x}_0, \dots, \mathbf{x}_{Ldt}\}$  in the inner variational cost to minimize. Ensemble members are generated from the Gaussian prior surrogate SPDE model, then conditioned by the neural solver so that the posterior pdf is no longer Gaussian. The posterior pdf is empirically ensemble-based computed:  $\mathbf{P}_{x|y} = \mathbf{X}\mathbf{X}^T$  with  $\mathbf{X} = (1/\sqrt{N-1}) \begin{bmatrix} \mathbf{x}_i^* - \mathbf{x}^* \end{bmatrix}$

472 Let note that in this formulation, the idea is to run SPDE-based conditional simulation of the prior  
 473 Matérn field  $X(\mathbf{s}, t)$ . One ensemble member is obtained by running one space-time simulation,  
 474 and two space-time neural reconstructions along the data assimilation window, then combined  
 475 through Eq. (21). As a consequence, there is no sequential assimilation. Though, in the idea,  
 476 such a conditioning is exactly similar to the one produced by EnKF simulations: in geostatistical  
 477 terms we can interpret the forecast step of the EnKF as being unconditional simulations at  
 478 time  $t - 1$  generating  $N$  realizations of a non-stationary random function (SPDE-based here)  
 479 for time  $t$ . Both prior mean and covariance matrix are then computed directly on this set of  
 480 unconditional realizations, before the analysis step, i.e. their conditioning with the observations.  
 481 The key advantage of the EnKF is its flow-dependency that propagates the uncertainties at each

482 time step with the evolution model, while the classic EnOI method generally used by geoscientists,  
483 see e.g. Asch et al. (2016); Counillon and Bertino (2009) replaces the flow-dependent EnKF  
484 error covariance matrix by a stationary matrix calculated from an historical ensemble. This is an  
485 important difference with our approach: while the neural architecture can be seen as a way to learn  
486 neural ensemble-based optimal interpolation models and solvers, our GP prior encoded by its  
487 SPDE precision matrix, built according to the FDM scheme with varying parameters over space  
488 and time, still allows for flow dependency, based on the learning of the SPDE parametrization  
489 given the input observations.

490

491 Such a strategy also shares many similarities with the so-called conditional generative models  
492 in deep learning, see (Goodfellow et al. 2014; Kingma and Welling 2022; Dinh et al. 2017; Ho  
493 et al. 2020). Instead of using neural networks to learn how to simulate in the prior distribution,  
494 we use here a generic class of advection-diffusion SPDE which we see as a first step towards  
495 physically-sounded generatives models in learning-based methods. Fig. 6 displays an example  
496 of such non-conditional simulations for the term  $\mathbf{x} - \mathbf{x}^b$ , again for Application 2 on realistic SSH  
497 datasets over some energetic area along the Gulf Stream, see Section 4, i.e. the anomaly between  
498 the true state sequence and the deterministic mean  $\mathbf{x}^b$ , retrieved from a preliminary 4DVarNet  
499 coarse resolution scheme. We can appreciate how both the generic class of SPDE selected here,  
500 together with the training of its parameter, leads to realistic anomalies with a clear increase of  
501 the variance along the main meander of the Gulf Stream, due to a correct distribution of SPDE  
502 parameter  $\tau^*(\mathbf{s}, t)$ . For a detailed analysis of the parameter estimation, please report to Section 4.

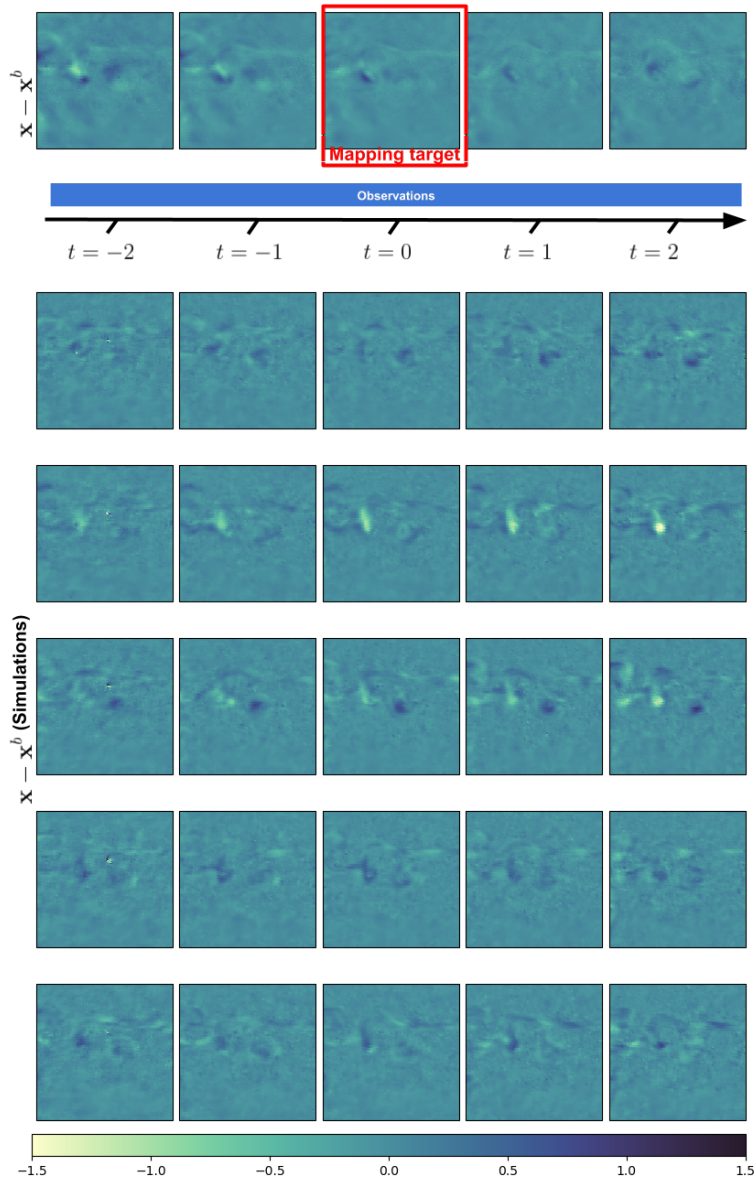


FIG. 6: Top panel: an example of the Ground Truth anomaly between  $\mathbf{x} - \mathbf{x}^b$ ; Bottom panel: five simulations of the same anomaly based on the SPDE parametrization  $\theta$  learned after training

#### 503 4. Results

504 In this Section, we provide two applications of this work:

- 505 • The first example relies on a spatio-temporal GP simulation driven by a non-stationary spatial  
506 diffusion tensor. Because the dynamical process is linear, the best reconstruction is provided

507 by the optimal interpolation using the SPDE parameters used in the simulation to fill in the  
 508 precision matrix.

- 509 • The second example uses an Observation System Simulation Experiment (OSSE) of the  
 510 Sea Surface Height (SSH) along the Gulf Stream. We will use the SPDE-based prior as a  
 511 surrogate model along the data assimilation window to provide ensemble members of the  
 512 posterior distribution.

513 *a. Diffusion-based non stationary GP*

514 **Dataset.** In this first application, we simulate 500 states of a GP driven by the following diffusion  
 515 SPDE:

$$\{\kappa^2 - \nabla \cdot \mathbf{H}(\mathbf{s})\nabla\}^{\alpha/2} \mathbf{x}(\mathbf{s}, t) = \tau \mathbf{z}(\mathbf{s}, t) \quad (22)$$

516 The regularity parameter  $\kappa = 0.33$  is fixed over space and time. To ensure the GP to be smooth  
 517 enough, we use a value of  $\alpha = 4$ . Such a formulation enables to generate GPs driven by local  
 518 anisotropies in space leading to non stationary spatio-temporal fields with eddy patterns. The  
 519 diffusion tensor  $\mathbf{H}$  is a 2-dimensional diffusion tensor generated by drawing from the spatial  
 520 statistics literature, see e.g. (Fuglstad et al. 2015a). We introduce a generic decomposition of  
 521  $\mathbf{H}(\mathbf{s}, t)$  through the equation:

$$\mathbf{H} = \gamma \mathbf{I}_2 + \beta \mathbf{v}(\mathbf{s})^T \mathbf{v}(\mathbf{s})$$

522 with  $\gamma = 1$ ,  $\beta = 25$  and  $\mathbf{v}(\mathbf{s}) = (v_1(\mathbf{s}), v_2(\mathbf{s}))^T$  using a periodic formulation of its two vector fields  
 523 components, see Section b. We use the Finite Difference Method in space coupled with an implicit  
 524 Euler scheme in time to solve the equation. Let  $\mathcal{D} = [0, 100] \times [0, 100]$  be the square spatial  
 525 domain of simulation and  $\mathcal{T} = [0, 500]$  the temporal domain. Both spatial and temporal domains  
 526 are discretized so that the simulation is made on a uniform Cartesian grid consisting of points  $(x_i,$   
 527  $y_j, t_k)$  where  $x_i = i\Delta x$ ,  $y_j = j\Delta y$ ,  $t_k = k\Delta t$  with  $\Delta x$ ,  $\Delta y$  and  $\Delta t$  all set to 1.

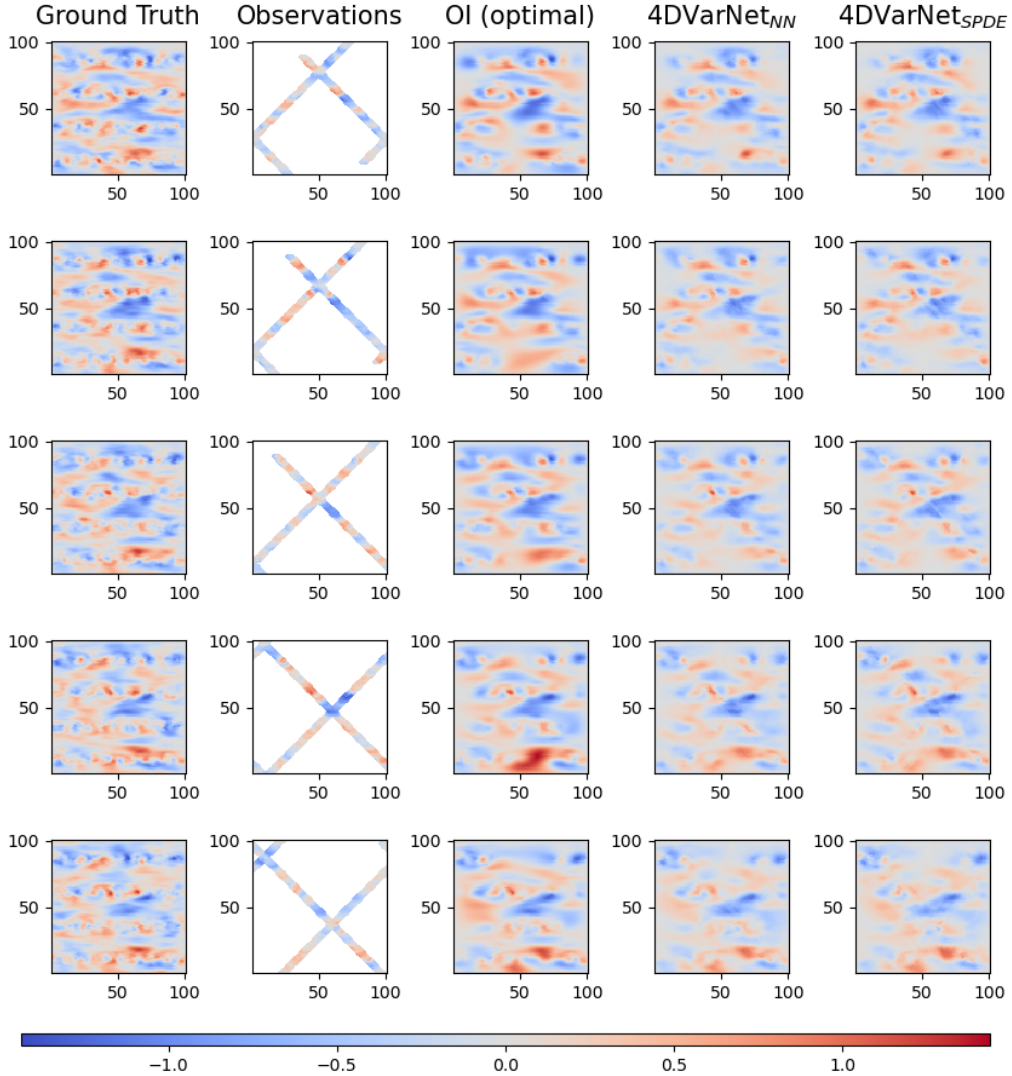


FIG. 7: From left to right: Ground Truth, Pseudo-observations, Optimal Interpolation, neural variational scheme with UNet-based and SPDE-based prior operator  $\Phi$ . A data assimilation window of length 5 is used

528 **Training setting.** We train the neural architectures for both UNet-based and SPDE-based priors  
 529 with Adam optimizer on 50 epochs. The training period goes from timestep 100 to 400. During  
 530 the training procedure, we select the best model according to metrics computed over the validation  
 531 period from timestep 30 to 80. Overall, the set of metrics is computed on a test period going from  
 532 timestep 400 to 500. No further improvements in the training losses are seen when training the

533 model longer. We use a data assimilation window of length 5 and generate pseudo-observations  
534 from the ground truth, inspired by orbiting satellites tracks around the earth, see Application 2 in  
535 Section b.

536  
537 **Results.** Fig. 7 displays the results obtained by Optimal Interpolation and the neural implicit  
538 solver with both UNet and SPDE-based parametrization of the prior. Neural-based reconstructions  
539 are optimized at the center of the assimilation window ( $T = 2$ ), which is why the performance  
540 may be affected for other leadtimes. There is no significant differences between the two prior  
541 formulations used in the neural scheme, which overall retrieve the main patterns of the OI. Some  
542 artefacts may appear due to the observation term in the inner variational cost that tends to lead the  
543 solution towards the observation in its close neighbourhood. Improvements may be expected when  
544 adding regularization terms in the training loss to counteract such effects, see e.g. Beauchamp  
545 et al. (2023b). Regarding the derived framework proposed in this work, one of the question was:  
546 is it possible to retrieve interpretable SPDE parametrizations from the joint learning setting? Two  
547 configurations were considered: when using as initial condition for the parametrization a gradient-  
548 based information of the accumulated alongtrack observations, i.e.  $\mathbf{H}_{11} = \nabla_{\vec{x}}\mathbf{y}$  and  $\mathbf{H}_{22} = \nabla_{\vec{y}}\mathbf{y}$ ; or  
549 an isotropic initialization, i.e.  $\mathbf{H} = \mathbf{I}$ . The first configuration enables to identify patterns in zonal  
550 and meridional components of the true diffusion tensor and leads to an optimal parametrization  
551  $\theta^*$  very close to the true diffusion tensor. While leading to different SPDE parametrizations  $\theta^*$ ,  
552 the interpolation metrics are similar in the end for the two initializations. In addition, using an  
553 isotropic initial condition is more general (see the next realistic SSH application) and also retrieves  
554 in the end the main zonal flow directions encoded by  $\mathbf{H}_{22}$ , while the meridional and periodic  
555 structures of  $\mathbf{H}_{11}$  and  $\mathbf{H}_{12}$  are partly seen as well.



556

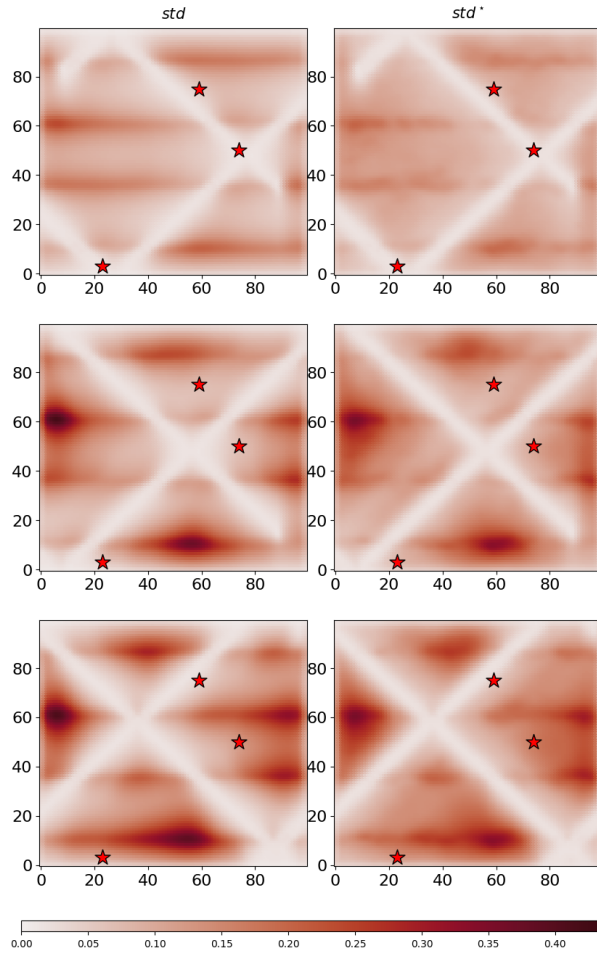


FIG. 8: True (left) and estimated (right) posterior standard deviations at the beginning, center and end of the assimilation window  $time = 0, 2, 4$

557

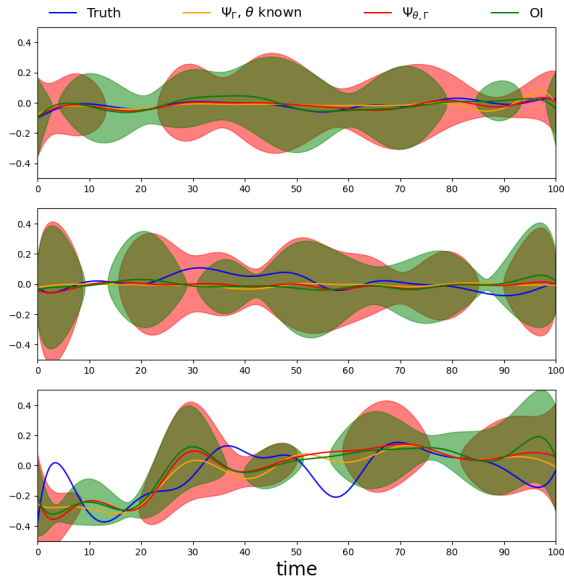


FIG. 9: Ground truth, OI and its posterior variance (blue), neural scheme with SPDE parametrization  $\theta$  known (orange) and with inference of  $\theta$  (red) for the same three points identified in Fig. 8 along the 100 time step test period

558 Because we know the process is Gaussian, we know that  $\mathbf{Q}(\mathbf{x}|\mathbf{y}) = \mathbf{Q}^b + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}$ , see Section  
559 2, to compute the closed form of the posterior pdf. Looking at the estimated posterior variance  
560 obtained starting from isotropic initial condition of the parameters, see Fig. 8, we understand that  
561 even if the initial set of parameters is not retrieved, they still remain interpretable. Visually, a sim-  
562 ulation produced by this set of SPDE parameters is also consistent and propose a spatio-temporal  
563 diffusion process close to the original one. It also highlights the dependency of the OI uncertainty  
564 quantification to the sampling scheme, especially here where no noise is added to the partial set  
565 of observations. On Fig. 9 is also shown the time series on three different locations (red stars)  
566  $\in \mathcal{D}$ , see again Fig. 8, of the GT (blue line) along the test period (100 time steps), the OI and  
567 corresponding standard deviation (green line), and the neural scheme estimations when the SPDE  
568 parametrization is known (orange line) or estimated (red line with associated uncertainty). Both  
569 neural scheme configurations are generally close which validates the capability of this framework  
570 to estimate jointly both state and prior parametrization. Because the neural scheme is optimized  
571 on the global MSE, its solution may deviate more or less significantly from the OI depending of  
572 the position we are looking at in domain  $\mathcal{D}$ .

573  
574 Last, Fig. 10 provides the scatterplot of the global MSE w.r.t the OI variational cost:

$$\mathcal{J}_{OI}(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}^*) = \|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2 + \mathbf{x}^T \mathbf{Q}_{\boldsymbol{\theta}^*}^b \mathbf{x}$$

575 throughout the iteration process after training of the neural schemes. For SPDE-based prior, the  
576 initial parametrization  $\boldsymbol{\theta}^{(0)}$  relates to an isotropic GP process. LSTM-based iterative solvers are all  
577 consistent with the optimal solution in terms of MSE. When the latter is used as training loss, 20  
578 iterations is enough to reach satisfactory performance. Using the same loss for inner variational  
579 cost and outer training loss (not shown here), see Beauchamp et al. (2022), would require more  
580 iterations to converge. Also, constraining the prior to follow the same SPDE simulation ensures  
581 to also jointly minimize the OI variational cost (asymptotic convergence of the red line to the  
582 yellow star on Fig. 10) which is not the case when looking for an optimal solution within the bi-  
583 level neural optimization of prior and solver (blue line) that may lead to deviate from the original  
584 variational cost to minimize. Let note that by construction, the analytical OI solution is optimal  
585 regarding the OI variational cost: it is unbiased with minimal variance. In other words, at a given

586 spatio-temporal location  $(s, t)$ , its variance (which is the local MSE) is minimal. In our case, we  
 587 compute the global MSE over the entire domain  $\mathcal{D} \times [t_k - 2, t_k + 2]$  w.r.t the true state because we  
 588 have only one single realization to compute this metrics. This is why the global MSE (see  $\mathcal{L}_1$ ) of  
 589 the OI may be outperformed by learning-based methods.

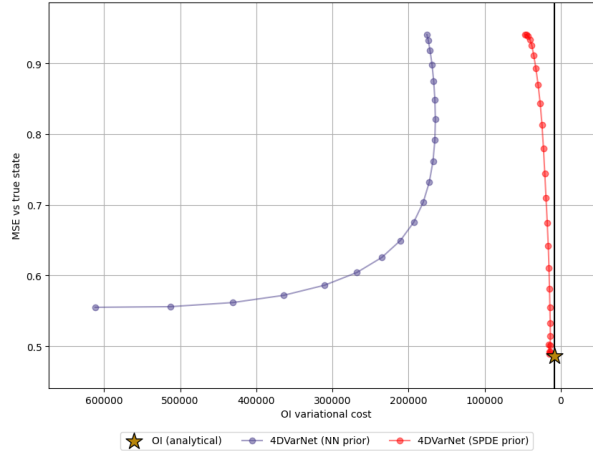


FIG. 10: Optimal Interpolation derived variational cost vs Mean Squared Error (MSE) loss (a) for the gradient-based descent of the variational cost, the classical implementation of the neural scheme and its SPDE-based prior formulation. For the analytical Optimal Interpolation solution (the yellow star), there is no iterations, then a single point is displayed.

590 *b. Realistic SSH datasets*

591 **Dataset.** In this application on Sea Surface Height (SSH) spatio-temporal fields, we focus on a  
 592 small part of the GULFSTREAM, see Fig. 11, mainly driven by energetic mesoscale dynamics,  
 593 to illustrate how our framework may help to solve for the oversmoothing of the state-of-the-art  
 594 Optimal interpolation (OI) and how the SPDE formulation of the prior is a consistent linearization  
 595 of the dynamics in the data assimilation window that helps to generate ensemble members in the  
 596 posterior distribution. We use an Observation System Simulation Experiment (OSSE) with the  
 597 NEMO (Nucleus for European Modeling of the Ocean) model NATL60 high resolution basin-  
 598 scale configuration Molines (2018). Based on this one-year long simulation, we generate pseudo  
 599 along-track nadir data for the current capabilities of the observation system (Ballarotta et al. 2019)  
 600 and pseudo wide-swath SWOT data in the context of the upcoming SWOT mission (Metref et al.  
 601 2020), with additional observation errors (Dufau et al. 2016; Esteban-Fernandez 2014; Gaultier  
 602 and Ubelmann 2010). The two types of observations may be merged, see Fig. 11, to produce a

603 one-year long daily datasets of partial and noisy observations of the idealized Ground Truth (GT).  
604 Last the DUACS operational system (CMEMS/C3S Copernicus program) provides the Optimal  
605 Interpolation baseline (Taburet et al. 2019) as daily gridded ( $0.25^\circ \times 0.25^\circ$ ) products

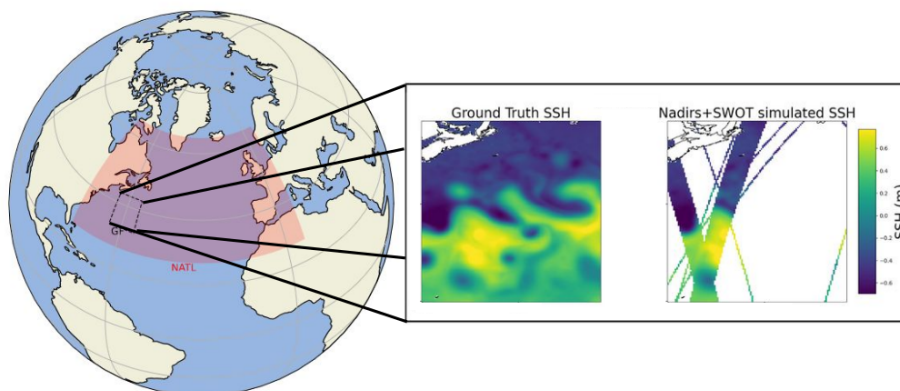


FIG. 11: NATL60 and GULFSTREAM domain and zoom-in picture for one day Ground Truth and accumulated along-track nadir + wide-swath SWOT SSH pseudo-observations

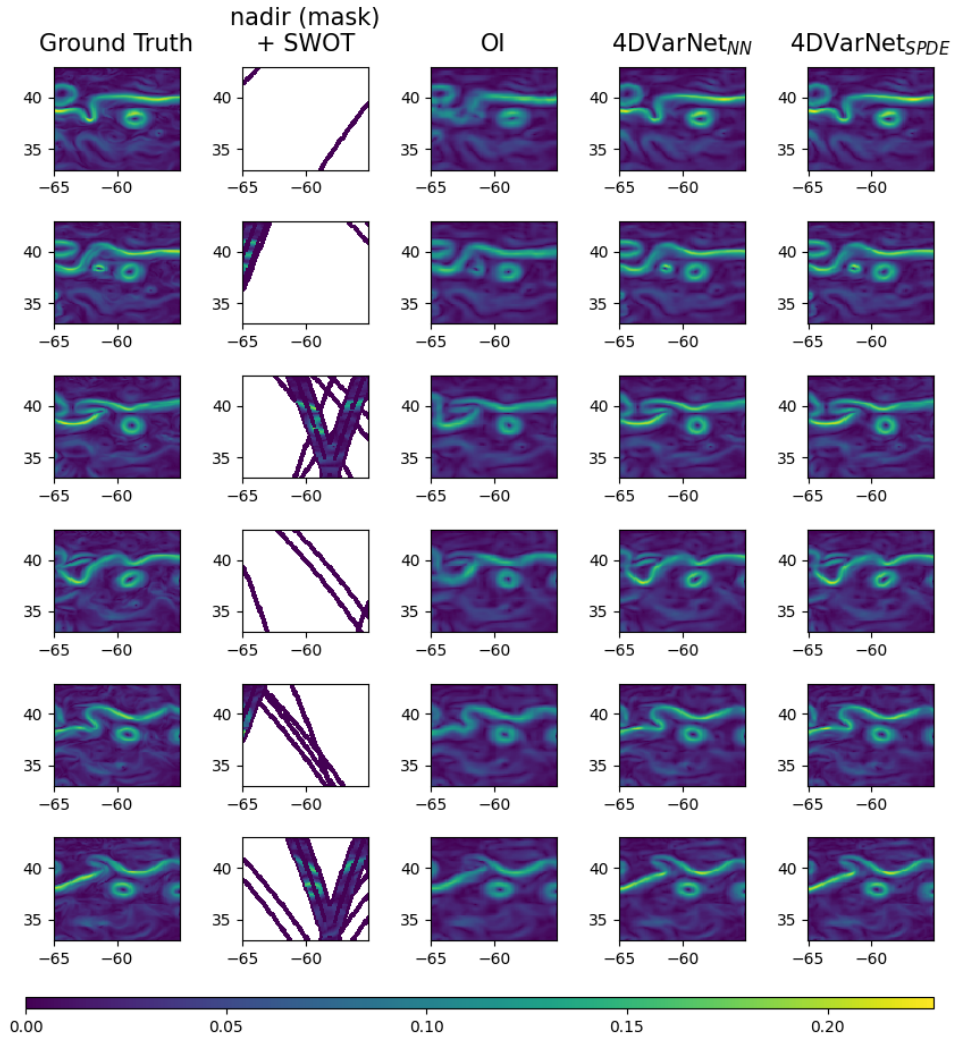


FIG. 12: From left to right: SSH Gradient Ground Truth, Pseudo-observations (nadir mask and SWOT Gradient field), DUACS Optimal Interpolation, and neural variational scheme results obtained with UNet-based and SPDE-based prior operator  $\Phi$ . A data assimilation window of length 5 is used

606 **Training setting.** All the datasets are downsampled from the original resolution of  $1/60^\circ$  to  $1/10^\circ$ .  
 607 For the training, the dataset spans from mid-February 2013 to October 2013, while the validation  
 608 period refers to January 2013. All methods are tested on the test period from October 22, 2012  
 609 to December 2, 2012. We still use Adam optimizer with 1100 epochs. Regarding the metrics, we  
 610 use the ocean data challenge <sup>1</sup> strategy looking at RMSE-score, and both spatial and temporel

<sup>1</sup>[https://github.com/ocean-data-challenges/2020a\\_SSH\\_mapping\\_NATL60](https://github.com/ocean-data-challenges/2020a_SSH_mapping_NATL60)

611 minimal scales resolved. The reader may refer to (Beauchamp et al. 2023b) for more details.

612

613 **Results.** Because we aim at assessing how the use of SPDE priors in the neural architecture is  
 614 relevant, we used a Gaussian prior formulation  $\mathbf{x} \sim \mathcal{N}(\mathbf{x}^b, \mathbf{Q}_\theta^b)$  where the mean  $\mathbf{x}^b$  is provided as a  
 615 first guess by a preliminary 4DVarNet run. Doing so, the SPDE parametrization  $\theta$  is focused on the  
 616 reconstruction and surrogate parametrization of the small scales not caught by this deterministic  
 617 mean. Then, we can reduce the data assimilation window used to estimate the prior mean  $\mathbf{x}^b$ , to a  
 618 reasonable length  $L = 5$  here, so that the storage of multiple (sparse) precision matrices throughout  
 619 the computational graph remains possible, see Section 3.d. Moving to longer time windows  
 620 including mesoscale-related autocorrelations (more than 10 days) in this SPDE framework would  
 621 lead to similar results, see e.g. Febvre et al. (2022) but would require moving to matrix-free  
 622 formulations, with potential existing solutions, see e.g. Pereira et al. (2022).

623 Last point on this experimental configuration: because the pseudo-observations are subsampled  
 624 from hourly simulations but we target daily reconstructions, they are noisy due to representativity  
 625 errors between the two temporal resolutions. This is not currently addressed by our framework where  
 626 the observation term in the minimization cost, see Eq. 8, is only the L2-norm of the innovations.  
 627 But it might be easily considered, either by using a known observation error covariance matrix  $\mathbf{R}$   
 628 or by learning one of its possible parametrization as an additional feature of the neural scheme.

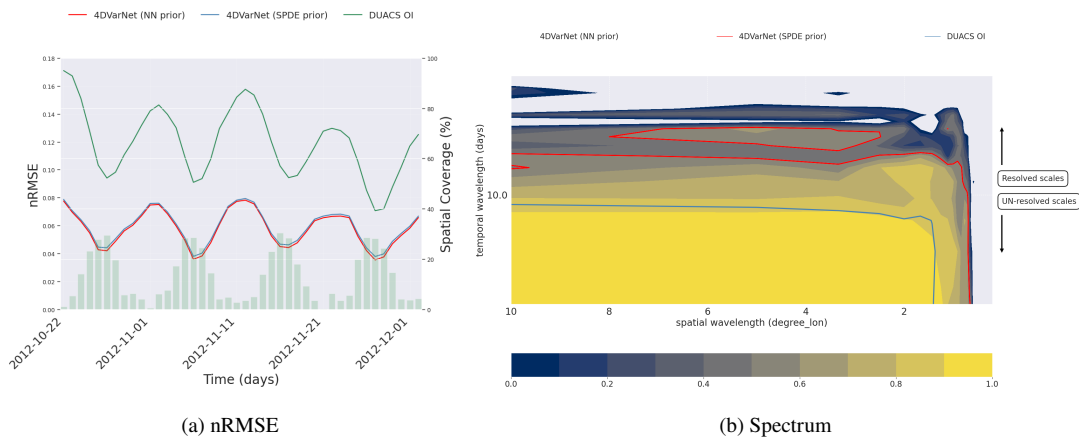


FIG. 13: For DUACS Optimal Interpolation, neural solvers with UNet-based and SPDE-based prior, (a) provides their temporal performance, i.e. nRMSE time series along the BOOST-SWOT DC evaluation period ; and (b) displays their spectral performance, i.e. the PSD-based score is used to evaluate the spatio-temporal scales resolved in the GULFSTREAM domain (yellow area)

629 As in Example 7.1, we provide in Fig. 12 the reconstructions, as the SSH gradients, obtained  
630 from DUACS OI baseline, and both neural solver formulations, the one using as prior a UNet-based  
631 parametrization, see e.g. (Beauchamp et al. 2023c) and our SPDE-based formulation. As expected  
632 and already seen in previous related studies (Beauchamp et al. 2023b,a; Fablet et al. 2021), the  
633 neural schemes improve the baseline by retrieving the dynamics along the main meander of the Gulf  
634 Stream and additional small energetic eddies. Again, there is no significant differences between  
635 the two neural formulations, which was again expected because the aim of the SPDE formulation  
636 is not to improve the mean state estimation obtained when using a neural prior operator, which  
637 is even more general, but to provide a a stochastic framework for interpretability and uncertainty  
638 quantification. This is supported by Figs. 13a and 13b respectively showing the normalized RMSE  
639 and the space-time spectrum along the test period. The periodic improvements of the score are  
640 due to the SWOT sampling that does not provide informations every day on this Gulf Stream  
641 domain. Overall, the nRMSE is in average improved by 60% when using the neural architecture.  
642 For the spectrum, minimal spatio-temporal scales  $\lambda_x$  and  $\lambda_t$  also improve resp. by 30% and 60%.  
643 From both figures and scores provided in Table 1, we can see that the UNet formulation of prior  
644  $\Phi$  leads to a small improvement in the reconstruction, which was expected because this is the  
645 only task optimized by the pure neural formulation. Introducing the SPDE formulation leads to  
646 optimize both reconstruction and likelihood of the parameters, which is more difficult. Though, the  
647 reconstruction performed by the latter is satisfactory enough and very close to the original solution  
648 proposed in Fablet et al. (2021). It also competes with other state-of-the-art method available in the  
649 ocean data challenges 2020a, among which DUACS OI (Taburet et al. 2019), MIOST (Multi-scale  
650 OI) Arduin et al. (2020) or a 4DVar scheme based on a QG dynamical model Le Guillou (2022).  
651 While DUACS OI has minimal spatial and temporal resolution of  $1.22^\circ$  and 11.15 days, 4DVarNet  
652 reaches  $0.62^\circ$  and 4.35 days, which reaches similar order of performance that combining a 4DVar  
653 with a QG model.

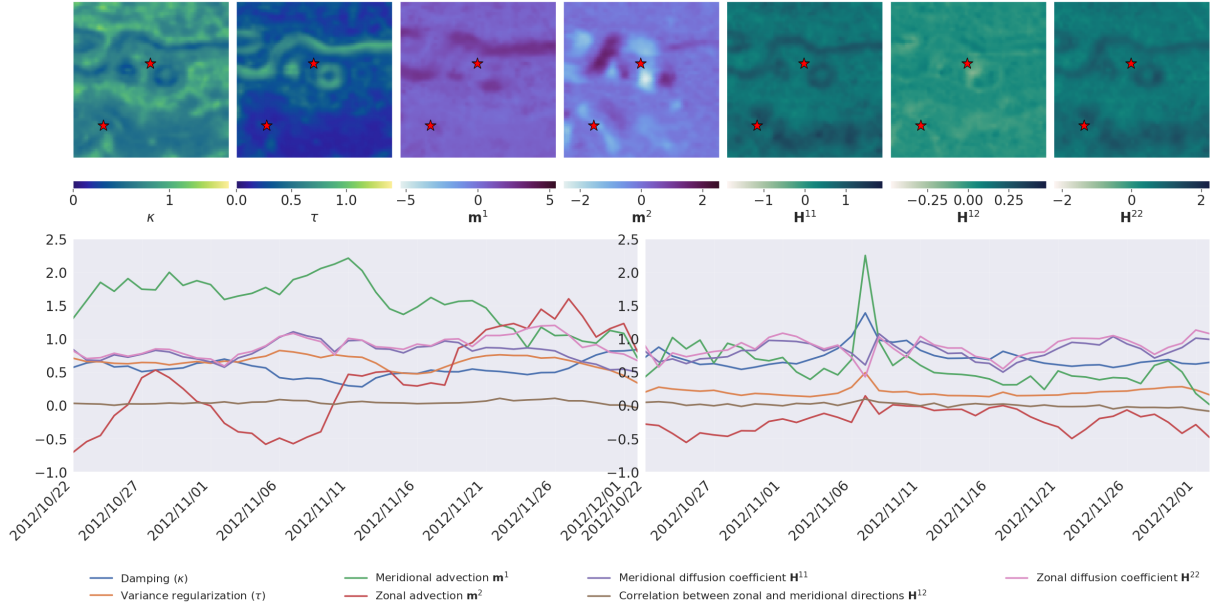


FIG. 14: Parameter estimation of the SPDE prior: in the neural scheme along the 42 days test period and every 6 days. Top, from left to right:  $\tau$ ,  $\mathbf{m}^1$ ,  $\mathbf{m}^2$  (advection fields),  $\mathbf{H}^{1,1}$ ,  $\mathbf{H}^{1,2}$  and  $\mathbf{H}^{2,2}$  (diffusion tensor) estimated on the first day of the test period. Bottom: time series of these parameters along the test period on the two locations identified by the red stars. The first one (left) is located along the Gulf Stream meander and the second one (right) is in the less energetic left-lower part of the domain.

654 In this very general setup, where the equation-based dataset provide a supervised learning setting  
655 on state  $\mathbf{x}$ , but not on the SPDE parameters  $\theta$ , two main questions are raised. On the first one:  
656 does the parameters retrieved by the iterative solver are interpretable? If considering Fig. 14 that  
657 shows parameters  $\tau$ ,  $\mathbf{m}_1$ ,  $\mathbf{m}_2$  (advection fields),  $\mathbf{H}_{11}$ ,  $\mathbf{H}_{12}$  and  $\mathbf{H}_{22}$  (diffusion tensor) on the first  
658 day of the test period, they seem consistent with the SSH field  $\mathbf{x}$  that partially encodes the SPDE  
659 parametrization, which also opens avenue for state-dependent parameters. We also show the time  
660 series along the 42 days of the test period for these parameters on two locations of the Gulf Stream  
661 domain (red stars): a first one right in the Gulf Stream meander (left time series) and a second  
662 one in the left-lower part of the domain, with less variability (right time series). Interestingly,  
663 in less energetic areas the parameters are almost perfectly correlated while in the Gulf Stream,  
664 they might behave differently. Playing with both damping and variance regularization parameters  
665 provide a flexible way to handle complex GP priors with both low and high marginal variances  
666 for a given time. This is a key aspect here because the range of possible values attributed to the  
667 anomaly generally differs according to the spatio-temporal dynamics of the SSH: it is high along



668 the main meander of the Gulf Stream and eddies not caught up by the OI, and lower elsewhere.  
669 This is also supported by generating non-conditional simulations based on these parameters and  
670 comparing it to the ground truth anomaly field, see again Fig. 6. The spatio-temporal fields are  
671 clearly consistent with the original simulation which also makes the link between our approach and  
672 generative modeling. The average of a large number of simulations would be a zero state, while its  
673 covariance relates to the model error matrix used in equation-based DA to perturb the members in  
674 ensemble methods.

	$\mu(\text{RMSE})$	$\sigma(\text{RMSE})$	$\lambda_x$ (degree)	$\lambda_t$ (days)
OI	0.92	0.02	1.22	11.06
MIOST	0.94	0.01	1.18	10.33
4DVar-QG	0.96	0.01	0.66	4.65
4DVarNet (NN prior)	0.96	0.01	0.62	4.35
4DVarNet (SPDE prior)	0.96	0.01	0.64	5.03

TABLE 1: Evaluation of the performance between OI, MIOST, 4DVar-QG, 4DVarNet schemes with UNet and SPDE-based prior parametrization. The OSSE involves 1 SWOT + 4 nadirs as pseudo-observations.

675 The second question is: does this approach, SPDE-based generation of members, followed by  
676 their conditioning with observations, is efficient to estimate the posterior pdf  $p_{\mathbf{x}|\mathbf{y}}$ ? Fig. 15 shows in  
677 a) the reconstruction error  $\mathbf{x} - \mathbf{x}^*$  for six days along the test period and in b) the empirical posterior  
678 standard deviations computed from 200 members. In c), we also provide pointwise Continuous  
679 Ranked Probability Score (CRPS) maps to assess the accuracy of the ensemble-based predictions.  
680 Given the observations  $\mathbf{y}_{ijk}$ , we compute the empirical CDF of the stochastic process  $X$  at a given  
681 spatio-temporal location  $\{i, j, k\}$  as  $F_{ijk}(z) = \mathbf{P}[X_{ijk} \leq z]$ :

$$CRPS(F_{ijk}, \mathbf{y}_{ijk}) = \int_{-\infty}^{+\infty} \left( F(z) - \mathbf{1}_{z - \mathbf{y}_{ijk}} \right)^2 dz$$

682 Looking at the estimated standard deviation produced by the ensemble of neural variational recon-  
683 structions, they are not as dependent of the observations as an OI scheme would be, which validates  
684 the flow-dependency discussed in Section e. We can still see the observation mask as blurry areas

685 but the standard deviations are rather increasing and continuous along the main meander of the  
 686 Gulf Stream. The CRPS is often close to zero, which indicates the ensemble of reconstruction is  
 687 wholly accurate with a realistic posterior standard deviation. The highest CRPS values observed  
 688 are about 0.4 which remains reasonable, and can be explained by high reconstruction errors outside  
 689 the main Gulf Stream meander that are not correctly handled by the ensemble.

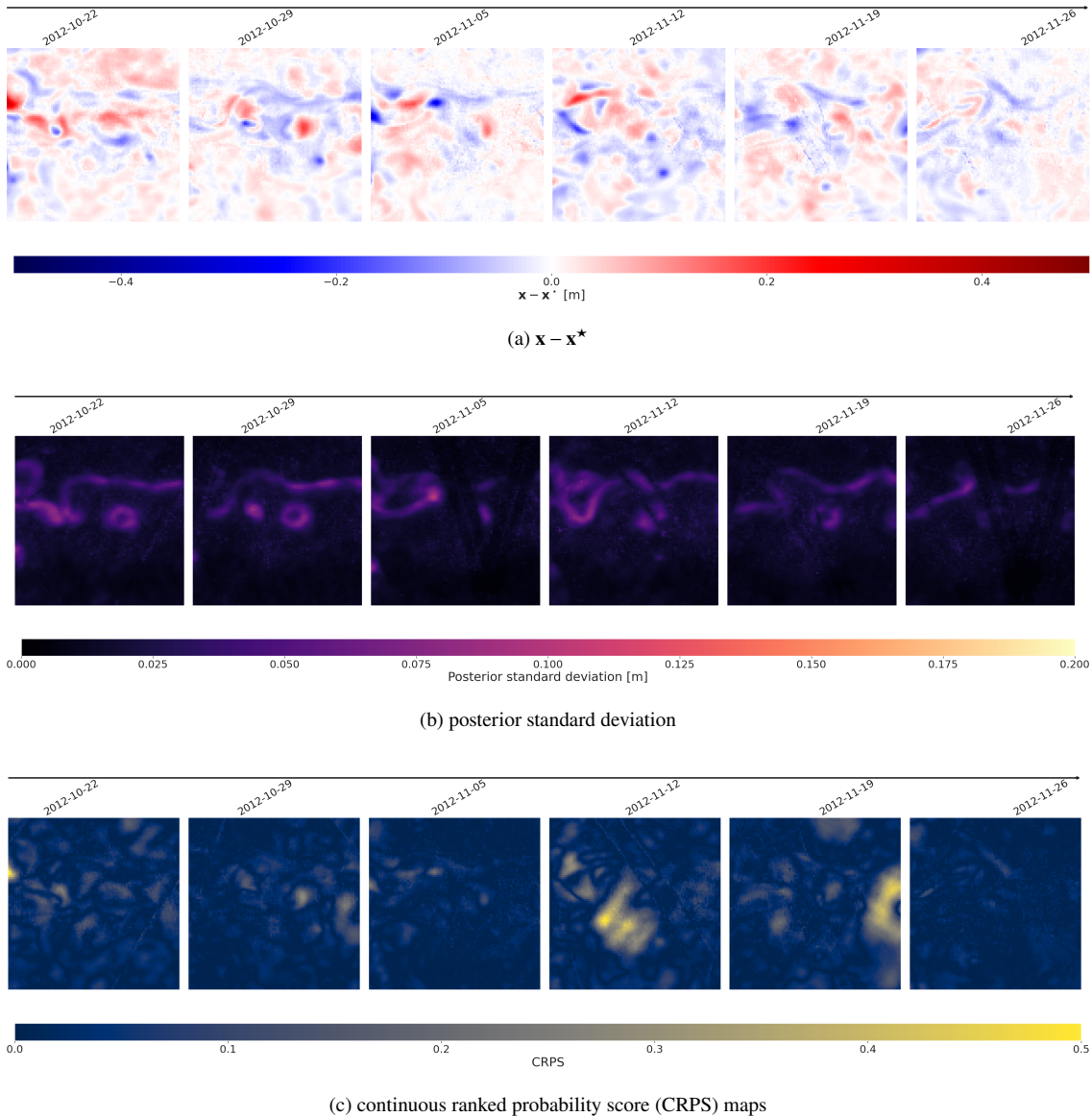


FIG. 15: For six days along the OSSE test period: a) corresponding reconstruction error at the center of the assimilation window, b) ensemble-based posterior standard deviations and c) Continuous Ranked Probability Score

## 690 5. Conclusion

691 We explore a new neural architecture to tackle the reconstruction inverse problem of a dynamical  
692 process from partial and potentially noisy observations. We provide a joint end-to-end learning  
693 scheme of both stochastic prior models and solvers. The idea is to optimize in the same time  
694 the state and the stochastic parametrization of the prior so that we minimize the mean squared  
695 error between the reconstruction and the true states, while we are also able to provide uncertainty  
696 quantification of the mean state reconstruction, either analytically in the gaussian case, or  
697 ensemble-derived in the more general configuration.

698  
699 In our work, we draw from recent advances in geostatistical simulations of SPDE-based Gaussian  
700 Processes to provide a flexible trainable prior embedded in a neural architecture backbone on  
701 variational data assimilation. The SPDE parameters are added as latent variables in an augmented  
702 state the trainable solver has to reconstruct. A bi-level optimization scheme is used to optimize in  
703 the same time:

- 704 • the inner variational cost derived from OI-based formulations, which depends on state  $\mathbf{x}$ ,  
705 observations  $\mathbf{y}$  and SPDE parameters  $\theta$  and,
- 706 • the outer training loss function of the neural architecture, which drives the optimization of  
707 the LSTM-based residual solver parameters  $\omega$  leading to the reconstruction of the augmented  
708 state.

709 The first application of the framework on a diffusion-based GP showed that it reaches the same  
710 performance, in terms of MSE w.r.t the ground truth, observed when using a neural-based prior  
711 in the classic implementation fo the neural scheme (Fablet et al. 2020), which asymptotically  
712 converges towards the optimal solution. In addition, the posterior variance of the mean state  
713 derived from the SPDE parameters is close to the true variance of the Optimal Interpolation,  
714 which demonstrates the potentiality of the proposed scheme to handle uncertainty quantifications.  
715 Indeed, if we showed that retrieving the original set of SPDE parameters might be difficult, the  
716 local minimum found after the training process leads to a parametrization with high likelihood,  
717 similar diffusion-based spatial patterns and spatio-temporal covariances structures. Though not  
718 always physically explainable, the SPDE prior formulation still helps to interpret the dynamical

719 process in terms of statistical properties.

720 In a more general setup, we also present an application on Sea Surface Height dynamics based  
721 on Observation System Simulation Experiment (OSSE), for which the ground truth is given  
722 by a state-of-the-art ocean model and pseudo-observations are generated by a realistic satellite  
723 subsampling of the ground truth. In this case, the process is not linear and the gaussian framework  
724 does not apply. Then, the idea is to use the GP linear SPDE formulation as a surrogate model to  
725 linearize the prior dynamics along the data assimilation window. It provides an efficient way for  
726 fast sampling of a huge set of members in the prior distribution within a few minutes. Based on  
727 the neural solver, able to handle non-linear and non-gaussian dynamics based on its supervised  
728 training, the conditioning of these simulations leads to the estimation of the posterior distribution.  
729 The preliminary conclusions made on the GP experiment holds for the reconstruction of the mean  
730 state, i.e. no significant differences observed when using a trainable neural prior. The key aspect  
731 of the framework is here revealed by the SPDE parametrization, which is fully non-stationary in  
732 space and time, and allows for online estimation after training for any new set of input observations,  
733 in contrast with other approaches in most of the spatial statistics literature which requires offline  
734 new parameter inference. We showed that the prior parametrization is both statistically consistent  
735 with the original ground truth used in the training and physically sounded with similar patterns  
736 observed in the simulations. Comparing the posterior pdf retrieved from the ensemble members,  
737 higher variances are observed in areas of high reconstruction errors which is a good indicator  
738 of the framework capability to quantify realistic uncertainties. Additional applications would be  
739 necessary on real datasets to complement these preliminary conclusions.

740

741 Regarding the potential extensions of this methodology, it is important to understand that such  
742 an SPDE parametrization of the prior term is a way of going to generative modelling, as it is  
743 called in the machine learning community. In our case, the SPDE is linear and already provide  
744 an efficient way to produce fast, large and realistic ensembles. Promising avenues would be to  
745 draw from the existing link between diffusion models and SDE to enrich our framework, provide  
746 stochastic non-linear priors and see if and how it helps to improve the results obtained in this work.  
747 Also, we made the choice to model the prior with a linear SPDE before conditioning it with our  
748 neural solver. A direct use of the SPDE formulation for the posterior would have been possible,

749 but restrictive to GP reconstructions which is in most cases rather limited. Back to the non-linear  
750 neural diffusion operators, an other way of addressing the problem would be to directly optimize  
751 the sampling in the posterior pdf.

752 *Acknowledgments.* This work has been supported by the LEFE program (LEFE MANU project  
753 IA-OAC), CNES (grant OSTST DUACS-HR), and ANR Projects Melody and OceaniX. This  
754 research has been supported by GENCI- IDRIS (grant no. 2020-101030).

755 *Data availability statement.* The open-source 4DVarNet with SPDE priors version of the  
756 code is available at <https://github.com/CIA-Oceanix/4dvarnet-starter/tree/maximebeauchamp>.  
757 The datasets are shared through the ocean data challenge 2020a also avail-  
758 able on GitHub [https://github.com/ocean-data-challenges/2020a\\\_SSH\\\_mapping\\\_NATL60](https://github.com/ocean-data-challenges/2020a\_SSH\_mapping\_NATL60),  
759 `lastaccess:2022`

### Finite difference schemes for SPDE

762 In this work, we propose to use the Euler implicit scheme as a discretization method for the  
 763 stochastic PDEs. We need to define the discretized version of several differential operators on the  
 764 2D regular grid, namely:

#### 765 a. Discretization of the spatial differential operators

$$\begin{aligned}
 \Delta \mathbf{x}_{i,j} &= \frac{\partial^2 \mathbf{x}}{\partial x^2}_{i,j} + \frac{\partial^2 \mathbf{x}}{\partial y^2}_{i,j} \\
 &= \frac{\partial}{\partial x} \left( \frac{\partial \mathbf{x}}{\partial x} \right)_{i,j} + \frac{\partial}{\partial y} \left( \frac{\partial \mathbf{x}}{\partial y} \right)_{i,j} \\
 &= \frac{\partial}{\partial x} \left( \frac{\mathbf{x}_{i+1,j} - \mathbf{x}_{i-1,j}}{2dx} \right) + \frac{\partial}{\partial y} \left( \frac{\mathbf{x}_{i,j+1} - \mathbf{x}_{i,j-1}}{2dy} \right) \\
 &= \left( \frac{\mathbf{x}_{i+1,j}}{dx^2} - \frac{\mathbf{x}_{i,j}}{dx^2} \right) - \left( \frac{\mathbf{x}_{i,j}}{dx^2} - \frac{\mathbf{x}_{i-1,j}}{dx^2} \right) \\
 &\quad + \left( \frac{\mathbf{x}_{i,j+1}}{dy^2} - \frac{\mathbf{x}_{i,j}}{dy^2} \right) - \left( \frac{\mathbf{x}_{i,j}}{dy^2} - \frac{\mathbf{x}_{i,j-1}}{dy^2} \right) \\
 &= \frac{\mathbf{x}_{i+1,j} - 2\mathbf{x}_{i,j} + \mathbf{x}_{i-1,j}}{dx^2} + \frac{\mathbf{x}_{i,j+1} - 2\mathbf{x}_{i,j} + \mathbf{x}_{i,j-1}}{dy^2}
 \end{aligned} \tag{A1a}$$

$$\begin{aligned}
\nabla \mathbf{H} \cdot \nabla \mathbf{x}_{i,j} &= \left( \mathbf{H}^{1,1} \frac{\partial}{\partial x^2} + \mathbf{H}^{1,2} \frac{\partial}{\partial x \partial y} + \mathbf{H}^{2,1} \frac{\partial}{\partial x \partial y} + \mathbf{H}^{2,2} \frac{\partial}{\partial y^2} \right) \mathbf{x}_{i,j} \\
&= \mathbf{H}^{1,1} \frac{\mathbf{x}_{i+1,j} - 2\mathbf{x}_{i,j} + \mathbf{x}_{i-1,j}}{dx^2} \\
&\quad + \mathbf{H}^{2,2} \frac{\mathbf{x}_{i,j+1} - 2\mathbf{x}_{i,j} + \mathbf{x}_{i,j-1}}{dy^2} \\
&\quad + \mathbf{H}^{1,2} \frac{\mathbf{x}_{i+1,j+1} - \mathbf{x}_{i+1,j-1} - \mathbf{x}_{i-1,j+1} + \mathbf{x}_{i-1,j-1}}{4dxdy} \\
&\quad + \mathbf{H}^{2,1} \frac{\mathbf{x}_{i+1,j+1} - \mathbf{x}_{i-1,j+1} - \mathbf{x}_{i+1,j-1} + \mathbf{x}_{i-1,j-1}}{4dxdy} \\
&= \mathbf{H}^{1,1} \frac{\mathbf{x}_{i+1,j} - 2\mathbf{x}_{i,j} + \mathbf{x}_{i-1,j}}{dx^2} \\
&\quad + \mathbf{H}^{2,2} \frac{\mathbf{x}_{i,j+1} - 2\mathbf{x}_{i,j} + \mathbf{x}_{i,j-1}}{dy^2} \\
&\quad + \mathbf{H}^{1,2} \frac{\mathbf{x}_{i+1,j+1} - \mathbf{x}_{i+1,j-1} - \mathbf{x}_{i-1,j+1} + \mathbf{x}_{i-1,j-1}}{2dxdy}
\end{aligned} \tag{A1b}$$

$$\begin{aligned}
\mathbf{m} \cdot \nabla \mathbf{x}_{i,j} &= \mathbf{m}^1 \frac{\partial}{\partial x} \mathbf{x}_{i,j} + \mathbf{m}^2 \frac{\partial}{\partial y} \mathbf{x}_{i,j} \\
&= \mathbf{m}^1 \frac{\mathbf{x}_{i+1,j} - \mathbf{x}_{i-1,j}}{2dx} + \mathbf{m}^2 \frac{\mathbf{x}_{i,j+1} - \mathbf{x}_{i,j-1}}{2dy}
\end{aligned} \tag{A1c}$$

768 *b. Upwind schemes for advection-dominated SPDE*

769 With such an advection-diffusion framework detailed in Section b, it is known that the solution to  
770 the space-centered scheme does not oscillate only when the Peclet number is lower than 2 and the  
771 Courant–Friedrichs–Lewy condition (CFL) condition  $Cr = dt (\mathbf{m}^1/dx + \mathbf{m}^2/dy) \leq 1$  is satisfied  
772 (Lewy et al. 1928; Price et al. 1966). For unsatisfied Peclet conditions, damped oscillations occur  
773 with nonreal eigenvalues (Finlayson 1992; Price et al. 1966), while in the limiting case of pure  
774 advection  $\mathbf{H} \rightarrow \mathbf{0}$ , such a scheme would be unconditionally unstable (Finlayson 1992; Strikwerda  
775 1989). Because the velocity field is allowed to vary in space and time, the CFL number is different  
776 at each discrete space-time location  $(i, j, t)$ . A necessary condition for convergence is that the CFL  
777 condition be satisfied at each point location, the velocity and diffusion parameters being unknown,

778 we have to choose the timestep  $dt$  small enough so that the maximum CFL number (observed in  
779 space at each time step) satisfies the CFL condition. Again, because the velocity field is trained, it  
780 may happen that the CFL condition is not satisfied if the latter is more and more dominant during  
781 the training. A simple way of counteracting this problem is to use an activation function on the  
782 two velocity components by clipping their maximum value.

783

784 One way of addressing this problem of stabilities in FDM when the advection term is predominant  
785 over the diffusion relates to the class of upwind schemes (UFDM). It is used to numerically simulate  
786 more properly the direction of propagation of the state in a flow field. The first order upwind FDM  
787 uses a one-sided finite difference in the upstream direction to approximate the advection term in  
788 the transport SPDE. The spatial accuracy of the first-order upwind scheme can be improved by  
789 choosing a more accurate finite difference stencil for the approximation of spatial derivative. Let  
790 note that UFDM scheme for SPDE eliminates the nonphysical oscillations in the space-centered  
791 scheme and generate stable solutions even for very complicated flows.

792 1) FIRST-ORDER UPWIND SCHEME (UFDM1)

793 Instead of using centered differences:

$$\left(\frac{\partial \mathbf{x}}{\partial x}\right)_{i,j} = \frac{\mathbf{x}_{i+1,j} - \mathbf{x}_{i-1,j}}{2dx} \quad (\text{A2a})$$

794

$$\left(\frac{\partial \mathbf{x}}{\partial y}\right)_{i,j} = \frac{\mathbf{x}_{i,j+1} - \mathbf{x}_{i,j-1}}{2dy} \quad (\text{A2b})$$

795 We use the one-sided upwind differences :

$$\left\{ \begin{array}{l} \left(\frac{\partial \mathbf{x}}{\partial x}\right)_{i,j} = \frac{\mathbf{x}_{i,j} - \mathbf{x}_{i-1,j}}{dx} \text{ if } \mathbf{m}_{i,j}^1 > 0 \\ \left(\frac{\partial \mathbf{x}}{\partial x}\right)_{i,j} = \frac{\mathbf{x}_{i+1,j} - \mathbf{x}_{i,j}}{dx} \text{ if } \mathbf{m}_{i,j}^1 < 0 \end{array} \right. \quad (\text{A3a})$$



796 and

$$\begin{cases} \left(\frac{\partial \mathbf{x}}{\partial y}\right)_{i,j} = \frac{\mathbf{x}_{i,j} - \mathbf{x}_{i,j-1}}{dy} \text{ if } \mathbf{m}_{i,j}^2 < 0 \\ \left(\frac{\partial \mathbf{x}}{\partial y}\right)_{i,j} = \frac{\mathbf{x}_{i,j+1} - \mathbf{x}_{i,j}}{dy} \text{ if } \mathbf{m}_{i,j}^2 > 0 \end{cases} \quad (\text{A3b})$$

797 2) THIRD-ORDER UPWIND SCHEME (UFDM3)

798 It can be shown, see e.g. (E. Ewing and Wang 2001) that the UFDM scheme is actually a second-  
 799 order approximation of the SPDE with a modified diffusion term. Along this line, it comes with  
 800 the family of methods that may introduce excessive numerical diffusion in the solution with large  
 801 gradients. Thus, we use a third order upwind scheme for the approximation of spatial derivatives  
 802 with four points instead of two, with only a reduced increase in the degree of sparsity of the  
 803 precision matrix. This scheme is less diffusive compared to the second-order accurate scheme. It  
 804 comes with four points instead of two for the approximation, with only a reduced increase in the  
 805 degree of sparsity of the discretized differential operator.

806 It can be expressed as follows:

$$\begin{cases} \left(\frac{\partial \mathbf{x}}{\partial x}\right)_{i,j} = \frac{2\mathbf{x}_{i+1,j} + 3\mathbf{x}_{i,j} - 6\mathbf{x}_{i-1,j} + \mathbf{x}_{i-2,j}}{6dx} \text{ if } \mathbf{m}_{i,j}^1 > 0 \\ \left(\frac{\partial \mathbf{x}}{\partial x}\right)_{i,j} = \frac{-\mathbf{x}_{i+2,j} + 6\mathbf{x}_{i+1,j} - 3\mathbf{x}_{i,j} - 2\mathbf{x}_{i-1,j}}{6dx} \text{ if } \mathbf{m}_{i,j}^1 < 0 \end{cases} \quad (\text{A4a})$$

807 and

$$\begin{cases} \left(\frac{\partial \mathbf{x}}{\partial y}\right)_{i,j} = \frac{2\mathbf{x}_{i,j+1} + 3\mathbf{x}_{i,j} - 6\mathbf{x}_{i,j-1} + \mathbf{x}_{i,j-2}}{6dy} \text{ if } \mathbf{m}_{i,j}^2 > 0 \\ \left(\frac{\partial \mathbf{x}}{\partial y}\right)_{i,j} = \frac{-\mathbf{x}_{i,j+2} + 6\mathbf{x}_{i,j+1} - 3\mathbf{x}_{i,j} - 2\mathbf{x}_{i,j-1}}{6dy} \text{ if } \mathbf{m}_{i,j}^2 < 0 \end{cases} \quad (\text{A4b})$$

808 *c. Discretization of the spatio-temporal SPDE*

809 Based on the discretization of the spatial operators using upwind finite difference schemes for  
 810 the advection term and centered difference schemes for the diffusion term, we involve an implicit

811 Euler scheme to solve the advection-diffusion SPDE:

$$\frac{\partial \mathbf{x}}{\partial t} + \left\{ \kappa^2(\mathbf{s}, t) + \mathbf{m}(\mathbf{s}, t) \cdot \nabla - \nabla \cdot \mathbf{H}(\mathbf{s}, t) \nabla \right\}^{\alpha/2} \mathbf{x}(\mathbf{s}, t) = \tau(\mathbf{s}, t) \mathbf{z}(\mathbf{s}, t)$$

812 Because the upwind advection schemes is not symmetric and does not involve the same neigh-  
 813 bours in the difference scheme approximation according to the predominant flow direction, we  
 814 introduce the following generic notations.

$$\begin{cases} \mathbf{m}_{i,j}^{1,t,-} = \left( \frac{\partial \mathbf{x}}{\partial x} \right)_{i,j}^t & \text{if } \mathbf{m}_{i,j}^{1,t} > 0 \\ \mathbf{m}_{i,j}^{1,t,+} = \left( \frac{\partial \mathbf{x}}{\partial x} \right)_{i,j}^t & \text{if } \mathbf{m}_{i,j}^{1,t} < 0 \end{cases} \quad \text{and} \quad \begin{cases} \mathbf{m}_{i,j}^{2,t,-} = \left( \frac{\partial \mathbf{x}}{\partial y} \right)_{i,j}^t & \text{if } \mathbf{m}_{i,j}^{2,t} > 0 \\ \mathbf{m}_{i,j}^{2,t,+} = \left( \frac{\partial \mathbf{x}}{\partial y} \right)_{i,j}^t & \text{if } \mathbf{m}_{i,j}^{2,t} < 0 \end{cases}$$

815 and by denoting  $\mathbf{a}_{i,j}^{1,t,+} = \max(\mathbf{m}_{i,j}^{1,t}, 0)$ ,  $\mathbf{a}_{i,j}^{1,t,-} = \min(\mathbf{m}_{i,j}^{1,t}, 0)$ ,  $\mathbf{a}_{i,j}^{2,t,+} = \max(\mathbf{m}_{i,j}^{2,t}, 0)$ ,  $\mathbf{a}_{i,j}^{2,t,-} =$   
 816  $\min(\mathbf{m}_{i,j}^{2,t}, 0)$ , the resulting UFD, whatever the order of the scheme, can be written in its compact  
 817 form as:

$$\begin{aligned} \mathbf{x}_{i,j}^{t+1} = & \mathbf{x}_{i,j}^t + dt \left[ \kappa_{i,j}^t \mathbf{x}_{i,j}^t + \left( \mathbf{a}_{i,j}^{1,t,+} \mathbf{m}_{i,j}^{1,t,-} + \mathbf{a}_{i,j}^{1,t,-} \mathbf{m}_{i,j}^{1,t,+} \right) + \left( \mathbf{a}_{i,j}^{2,t,+} \mathbf{m}_{i,j}^{2,t,-} + \mathbf{a}_{i,j}^{2,t,-} \mathbf{m}_{i,j}^{2,t,+} \right) \right. \\ & + \mathbf{H}_{i,j}^{1,1,t} \frac{\mathbf{x}_{i+1,j}^t - 2\mathbf{x}_{i,j}^t + \mathbf{x}_{i-1,j}^t}{dx^2} + \mathbf{H}_{i,j}^{2,2,t} \frac{\mathbf{x}_{i,j+1}^t - 2\mathbf{x}_{i,j}^t + \mathbf{x}_{i,j-1}^t}{dy^2} \\ & \left. + \mathbf{H}_{i,j}^{1,2,t} \frac{\mathbf{x}_{i+1,j+1}^t - \mathbf{x}_{i+1,j-1}^t - \mathbf{x}_{i-1,j+1}^t + \mathbf{x}_{i-1,j-1}^t}{2dxdy} + \tau_{i,j}^t \mathbf{z}_{i,j} \right] \end{aligned}$$

818 Using again notation  $i = \lfloor k/N_x \rfloor$  and  $j = k \bmod N_x$ , operator  $\mathbf{A}_t$  associated to UFD1 finally  
819 writes:

$$\mathbf{A}_{k,l}(t) = \begin{cases} \mathbf{H}_{i,j}^{1,2,t}/2dxdy & \text{if } l = k \pm (N_x + 1) \\ \mathbf{H}_{i,j}^{1,2,t}/2dxdy & \text{if } l = k \pm (N_x - 1) \\ -\mathbf{H}_{i,j}^{1,1,t}/dx^2 + (\mathbf{a}_{i,j}^{1,t,+} + \mathbf{a}_{i,j}^{1,t,-})/dx & \text{if } l = k - 1 \\ -\mathbf{H}_{i,j}^{1,1,t}/dx^2 + (\mathbf{a}_{i,j}^{1,t,+} + \mathbf{a}_{i,j}^{1,t,-})/dx & \text{if } l = k + 1 \\ -\mathbf{H}_{i,j}^{2,2,t}/dy^2 + (\mathbf{a}_{i,j}^{2,t,+} + \mathbf{a}_{i,j}^{2,t,-})/dy & \text{if } l = k - N_x \\ -\mathbf{H}_{i,j}^{2,2,t}/dy^2 + (\mathbf{a}_{i,j}^{2,t,+} + \mathbf{a}_{i,j}^{2,t,-})/dy & \text{if } l = k + N_x \\ (\kappa_{i,j}^t)^2 + 2(\mathbf{H}_{i,j}^{1,1,t}/dx^2 + \mathbf{H}_{i,j}^{2,2,t}/dy^2) \\ + (\mathbf{a}_{i,j}^{1,t,+} + \mathbf{a}_{i,j}^{1,t,-})/dx + (\mathbf{a}_{i,j}^{2,t,+} + \mathbf{a}_{i,j}^{2,t,-})/dy & \text{if } k = l \\ 0 & \text{otherwise} \end{cases} \quad (\text{A5})$$

820 The same operator associated to UFD3 writes:

$$\mathbf{A}_{k,l}(t) = \begin{cases} \mathbf{H}_{i,j}^{1,2,t}/2dxdy & \text{if } l = k \pm (N_x + 1) \\ \mathbf{H}_{i,j}^{1,2,t}/2dxdy & \text{if } l = k \pm (N_x - 1) \\ \mathbf{a}_{i,j}^{1,t,+}/6dx & \text{if } l = k - 2 \\ -\mathbf{H}_{i,j}^{1,1,t}/dx^2 - (6\mathbf{a}_{i,j}^{1,t,+} + 2\mathbf{a}_{i,j}^{1,t,-})/6dx & \text{if } l = k - 1 \\ -\mathbf{H}_{i,j}^{1,1,t}/dx^2 + (2\mathbf{a}_{i,j}^{1,t,+} + 6\mathbf{a}_{i,j}^{1,t,-})/6dx & \text{if } l = k + 1 \\ -\mathbf{a}_{i,j}^{1,t,-}/6dx & \text{if } l = k + 2 \\ \mathbf{a}_{i,j}^{2,t,+}/6dy & \text{if } l = k - 2N_x \\ -\mathbf{H}_{i,j}^{2,2,t}/dy^2 - (6\mathbf{a}_{i,j}^{2,t,+} + 2\mathbf{a}_{i,j}^{2,t,-})/6dy & \text{if } l = k - N_x \\ -\mathbf{H}_{i,j}^{2,2,t}/dy^2 + (2\mathbf{a}_{i,j}^{2,t,+} + 6\mathbf{a}_{i,j}^{2,t,-})/6dy & \text{if } l = k + N_x \\ -\mathbf{a}_{i,j}^{2,t,-}/6dy & \text{if } l = k + 2N_x \\ (\kappa_{i,j}^t)^2 + 2(\mathbf{H}_{i,j}^{1,1,t}/dx^2 + \mathbf{H}_{i,j}^{2,2,t}/dy^2) \\ + 3(\mathbf{a}_{i,j}^{1,t,+} - \mathbf{a}_{i,j}^{1,t,-})/6dx \\ + 3(\mathbf{a}_{i,j}^{2,t,+} - \mathbf{a}_{i,j}^{2,t,-})/6dy & \text{if } k = l \end{cases} \quad (\text{A6})$$

821 and 0 otherwise.

### SPDE-based precision matrix

824 Let define the spatio-temporal SPDE prior  $\mathbf{x} = \{\mathbf{x}_0, \dots, \mathbf{x}_{Ldt}\}$ . From now on,  $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{P}_0)$   
 825 denotes the initial state and  $\mathbf{Q}_0 = \mathbf{P}_0^{-1}$  is always taken as the precision matrix obtained after a given  
 826 stabilization run, i.e. the evolution of the dynamical system over  $N$  timesteps using as stationary  
 827 parameters the initial parametrization  $\theta_0$  of the SPDE at time  $t = 0$ . We can rewrite :

$$\{\mathbf{x}_0, \dots, \mathbf{x}_{Ldt}\} = \mathbf{M}_G \begin{bmatrix} \mathbf{x}_0 \\ \mathbf{z} \end{bmatrix}$$

828 with  $\mathbf{z} = [\mathbf{z}_1, \dots, \mathbf{z}_t]^T$  and

$$\mathbf{M}_G = \begin{bmatrix} \mathbf{I} & 0 & 0 & 0 & 0 & \dots & 0 \\ \mathbf{M}_1 & \mathbf{T}_1 & 0 & 0 & 0 & \dots & 0 \\ \mathbf{M}_2\mathbf{M}_1 & \mathbf{M}_2\mathbf{T}_1 & \mathbf{T}_2 & 0 & 0 & \dots & 0 \\ \mathbf{M}_3\mathbf{M}_2\mathbf{M}_1 & \mathbf{M}_3\mathbf{M}_2\mathbf{T}_1 & \mathbf{M}_3\mathbf{T}_2 & \mathbf{T}_3 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \mathbf{T}_L \end{bmatrix}$$

829 Despite its apparent complexity,  $\mathbf{M}_G$  has a particular structure which allows to easily compute  
 830 its inverse:

$$\mathbf{M}_G^{-1} = \begin{bmatrix} \mathbf{I} & 0 & 0 & 0 & 0 & \dots & 0 \\ -\mathbf{T}_1^{-1}\mathbf{M}_1 & \mathbf{T}_1^{-1} & 0 & 0 & 0 & \dots & 0 \\ 0 & -\mathbf{T}_2^{-1}\mathbf{M}_2 & \mathbf{T}_2^{-1} & 0 & 0 & \dots & 0 \\ 0 & 0 & -\mathbf{T}_3^{-1}\mathbf{M}_3 & \mathbf{T}_3^{-1} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & -\mathbf{T}_L^{-1}\mathbf{M}_L & \mathbf{T}_L^{-1} \end{bmatrix}$$

831 All this information is embedded in  $\mathbf{Q}^b$  (Eq. B2), which is the inverse of the prior covariance  
 832 matrix  $\mathbf{B}$ :

$$\mathbf{Q}^b = \mathbf{B}^{-1} = \begin{bmatrix} \mathbf{P}_0 & \mathbf{P}_{0,1} & \dots & \dots & \mathbf{P}_{0,L} \\ \mathbf{P}_{1,0} & \mathbf{P}_1 & \dots & \dots & \mathbf{P}_{1,L} \\ \vdots & \mathbf{P}_{2,1} & \mathbf{P}_2 & \dots & \mathbf{P}_{2,L} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{P}_{L-1,0} & \dots & \dots & \mathbf{P}_{L-1} & \mathbf{P}_{L-1,L} \\ \mathbf{P}_{L,0} & \mathbf{P}_{L,1} & \dots & \mathbf{P}_{L,L-1} & \mathbf{P}_L \end{bmatrix}^{-1} = \mathbf{M}_G^{-1\text{T}} \begin{bmatrix} \mathbf{P}_0^{-1} & 0 & \dots & 0 \\ 0 & \mathbf{I} & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{I} \end{bmatrix} \mathbf{M}_G^{-1} \quad (\text{B1})$$

833 By denoting  $\mathbf{S}_k = \mathbf{T}_k \mathbf{T}_k^{\text{T}}$ , we have

$$\mathbf{Q}^b = \begin{bmatrix} \mathbf{P}_0^{-1} + \mathbf{M}_1^{\text{T}} \mathbf{S}_1^{-1} \mathbf{M}_1 & -\mathbf{M}_1^{\text{T}} \mathbf{S}_1^{-1} & 0 & 0 & 0 & \dots & 0 \\ -\mathbf{S}_1^{-1} \mathbf{M}_1 & \mathbf{S}_1^{-1} + \mathbf{M}_2^{\text{T}} \mathbf{S}_2^{-1} \mathbf{M}_2 & -\mathbf{M}_2^{\text{T}} \mathbf{S}_2^{-1} & 0 & 0 & \dots & 0 \\ 0 & -\mathbf{S}_2^{-1} \mathbf{M}_2 & \mathbf{S}_2^{-1} + \mathbf{M}_3^{\text{T}} \mathbf{S}_3^{-1} \mathbf{M}_3 & -\mathbf{M}_3^{\text{T}} \mathbf{S}_3^{-1} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -\mathbf{S}_{L-1}^{-1} \mathbf{M}_{L-1} & \mathbf{S}_{L-1}^{-1} + \mathbf{M}_L^{\text{T}} \mathbf{S}_L^{-1} \mathbf{M}_L & -\mathbf{M}_L^{\text{T}} \mathbf{S}_L^{-1} \\ 0 & \ddots & \ddots & \ddots & 0 & -\mathbf{S}_L^{-1} \mathbf{M}_L & \mathbf{S}_L^{-1} \end{bmatrix} \quad (\text{B2})$$

834 Because of the formulation of  $\mathbf{M}_t$  and  $\mathbf{T}_t$ , the precision matrix  $\mathbf{Q}^b$  with the FDM scheme also  
 835 writes:

$$\mathbf{Q}^b = \frac{1}{dt} \begin{bmatrix} \mathbf{P}_0^{-1} + \tilde{\mathbf{Q}}_{s,1} & -\tilde{\mathbf{Q}}_{s,1} \mathbf{M}_1^{-1} & 0 & 0 & 0 & \dots & 0 \\ -(\mathbf{M}_1^{\text{T}})^{-1} \tilde{\mathbf{Q}}_{s,1} & \mathbf{M}_1^{\text{T}} \tilde{\mathbf{Q}}_{s,1} \mathbf{M}_1 + \tilde{\mathbf{Q}}_{s,2} & -\tilde{\mathbf{Q}}_{s,2} \mathbf{M}_2^{-1} & 0 & 0 & \dots & 0 \\ 0 & -(\mathbf{M}_2^{\text{T}})^{-1} \tilde{\mathbf{Q}}_{s,2} & \mathbf{M}_2^{\text{T}} \tilde{\mathbf{Q}}_{s,2} \mathbf{M}_2 + \tilde{\mathbf{Q}}_{s,3} & -\tilde{\mathbf{Q}}_{s,3} \mathbf{M}_3^{-1} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -(\mathbf{M}_{L-1}^{\text{T}})^{-1} \tilde{\mathbf{Q}}_{s,L-1} & \mathbf{M}_{L-1}^{\text{T}} \tilde{\mathbf{Q}}_{s,L-1} \mathbf{M}_{L-1} + \tilde{\mathbf{Q}}_{s,L} & -\tilde{\mathbf{Q}}_{s,L} \mathbf{M}_L^{-1} \\ 0 & \ddots & \ddots & \ddots & 0 & -(\mathbf{M}_L^{\text{T}})^{-1} \tilde{\mathbf{Q}}_{s,L} & \mathbf{M}_L^{\text{T}} \tilde{\mathbf{Q}}_{s,L} \mathbf{M}_L \end{bmatrix} \quad (\text{B3})$$

836 where  $\tilde{\mathbf{Q}}_{s,t}$  is the precision matrix of the colored noise weighted by the non-uniform regularization  
 837 variance  $\tau_t$ .

838

840

### PyTorch implementation of sparse linear algebra

841

842

Currently, two pieces of codes are missing in the PyTorch sparse linear algebra to achieve a fully sparse implementation of our algorithm: the automatic differentiation tools for

843

- solving sparse linear systems

844

- running sparse Cholesky decomposition

845

846

First, regarding the implementation of the backward pass for solving linear systems, we start by writing the forward pass of this system of equation:

$$x = \mathbf{A}^{-1}\mathbf{b}$$

847

where  $\mathbf{A}$  denotes a  $2D$  square matrix and  $\mathbf{b}$  a one-dimensional vector.

848

We need to provide the gradients wrt both  $\mathbf{A}$  and  $\mathbf{b}$ :

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{b}} &= \frac{\partial L}{\partial x_i} \frac{\partial x_i}{\partial \mathbf{b}_j} = \frac{\partial L}{\partial x_i} \frac{\partial}{\partial \mathbf{b}_k} (\mathbf{A}_{ij}^{-1} \mathbf{b}_j) = \frac{\partial L}{\partial x_i} \mathbf{A}_{ij}^{-1} \frac{\partial \mathbf{b}_j}{\partial \mathbf{b}_k} \\ &= \frac{\partial L}{\partial x_i} \mathbf{A}_{ij}^{-1} \delta_{jk} = \frac{\partial L}{\partial x_i} \mathbf{A}_{ik}^{-1} = (\mathbf{A}^{-1})^T \frac{\partial L}{\partial x} \\ &= \text{solve}(\mathbf{A}^T, \frac{\partial L}{\partial x}) \end{aligned}$$

849

and

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{A}} &= \frac{\partial L}{\partial x_i} \frac{\partial x_i}{\partial \mathbf{A}_{mn}} = \frac{\partial L}{\partial x_i} \frac{\partial}{\partial \mathbf{A}_{mn}} (\mathbf{A}_{ij}^{-1} \mathbf{b}_j) \\
&= -\frac{\partial L}{\partial x_i} \mathbf{A}_{ij}^{-1} \frac{\partial \mathbf{A}_{jk}}{\partial \mathbf{A}_{mn}} \mathbf{A}_{kl}^{-1} \mathbf{b}_l \\
&= -\frac{\partial L}{\partial x_i} \mathbf{A}_{ij}^{-1} \delta_{jm} \delta_{kn} \mathbf{A}_{kl}^{-1} \mathbf{b}_l \\
&= -\frac{\partial L}{\partial x_i} \mathbf{A}_{im}^{-1} \mathbf{A}_{nl}^{-1} \mathbf{b}_l \\
&= -\left( (\mathbf{A}^{-1})^T \frac{\partial L}{\partial x} \right) \otimes (\mathbf{A}^{-1} \mathbf{b}) \\
&= -\frac{\partial L}{\partial \mathbf{b}} \otimes x
\end{aligned}$$

850 where we used the einstein summation convention during the derivations, as well as the following  
851 identity:

$$\frac{\partial (\mathbf{A}^{-1})}{\partial p} = -\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial p} \mathbf{A}^{-1}.$$

852 These two expressions  $\frac{\partial L}{\partial \mathbf{b}}$  and  $\frac{\partial L}{\partial \mathbf{A}}$  are easily implemented in PyTorch based on sparse representa-  
853 tions.

854  
855 Second, the backward pass for a sparse Cholesky decomposition may be found for instance in  
856 Seeger et al. (2019): given a symmetric, positive definite matrix  $\mathbf{A}$ , its Cholesky factor  $\mathbf{L}$  is lower  
857 triangular with positive diagonal, such that the forward pass of the Cholesky decomposition is  
858 defined as  $\mathbf{A} = \mathbf{L}\mathbf{L}^T$ . Given the output gradient  $\bar{\mathbf{L}}$  and the Cholesky factor  $\mathbf{L}$ , the backward pass  
859 compute the input gradient  $\bar{\mathbf{A}}$  defined as :

$$\bar{\mathbf{A}} = \frac{1}{2} \mathbf{L}^{-T} \text{ltu}(\mathbf{L}^T \bar{\mathbf{L}}) \mathbf{L}^{-1} \tag{C1}$$

860 where  $\text{ltu}(\mathbf{X})$  generates a symmetric matrix by copying the lower triangle to the upper triangle.  
861 In our work, such an operation is useful when optimizing the likelihood in the outer training cost  
862 function of the neural scheme, see Eq. 20, which involves the determinant of sparse matrices  
863 Cholesky decomposition.

## 864 **References**

- 865 M. Andrychowicz, M. Denil, S. Gomez, M. W. Hoffman, D. Pfau, T. Schaul, B. Shillingford, and  
866 N. De Freitas. Learning to learn by gradient descent by gradient descent. In *Advances in neural*  
867 *information processing systems*, pages 3981–3989, 2016.
- 868 F. Ardhuin, C. Ubelmann, G. Dibarboure, L. Gaultier, A. Ponte, M. Ballarotta, and Y. Faugère.  
869 Reconstructing ocean surface current combining altimetry and future spaceborne doppler data.  
870 *Earth and Space Science Open Archive*, page 22, 2020. doi: 10.1002/essoar.10505014.1. URL  
871 <https://doi.org/10.1002/essoar.10505014.1>.
- 872 L. Arras, J. Arjona-Medina, M. Widrich, G. Montavon, M. Gillhofer, K.-R. Müller, S. Hochreiter,  
873 and W. Samek. Explaining and interpreting lstms. *Lecture Notes in Computer Science*, pages  
874 211–238, 2019. ISSN 1611-3349. doi: 10.1007/978-3-030-28954-6\_11. URL [http://dx.](http://dx.doi.org/10.1007/978-3-030-28954-6_11)  
875 [doi.org/10.1007/978-3-030-28954-6\\_11](http://dx.doi.org/10.1007/978-3-030-28954-6_11).
- 876 M. Asch, M. Bocquet, and M. Nodet. *Data Assimilation. Fundamentals of Algorithms*. Society  
877 for Industrial and Applied Mathematics, Dec. 2016. ISBN 978-1-61197-453-9. doi: 10.1137/1.  
878 9781611974546. URL <https://doi.org/10.1137/1.9781611974546>.
- 879 M. Ballarotta, C. Ubelmann, M.-I. Pujol, G. Taburet, F. Fournier, J.-F. Legeais, Y. Faugère,  
880 A. Delepouille, D. Chelton, G. Dibarboure, and N. Picot. On the resolutions of ocean altimetry  
881 maps. *Ocean Science*, 15(4):1091–1109, 2019. doi: 10.5194/os-15-1091-2019. URL [https:](https://www.ocean-sci.net/15/1091/2019/)  
882 [//www.ocean-sci.net/15/1091/2019/](https://www.ocean-sci.net/15/1091/2019/).
- 883 M. Beauchamp, J. Thompson, H. Georgenthum, Q. Febvre, and R. Fablet. Learning neural optimal  
884 interpolation models and solvers, 2022. URL <https://arxiv.org/abs/2211.07209>.
- 885 M. Beauchamp, Q. Febvre, and R. Fablet. Ensemble-based 4dvarnet uncertainty quantification for  
886 the reconstruction of sea surface height dynamics. *Environmental Data Science*, 2:e18, 2023a.  
887 doi: 10.1017/eds.2023.19.
- 888 M. Beauchamp, Q. Febvre, H. Georgenthum, and R. Fablet. 4dvarnet-ssh: end-to-end learning  
889 of variational interpolation schemes for nadir and wide-swath satellite altimetry. *Geoscientific*  
890 *Model Development*, 16(8):2119–2147, 2023b. doi: 10.5194/gmd-16-2119-2023. URL [https:](https://gmd.copernicus.org/articles/16/2119/2023/)  
891 [//gmd.copernicus.org/articles/16/2119/2023/](https://gmd.copernicus.org/articles/16/2119/2023/).



- 892 M. Beauchamp, Q. Febvre, J. Thompson, H. Georgenthum, and R. Fablet. Learning neural  
893 optimal interpolation models and solvers. In J. Mikyška, C. de Mulatier, M. Paszynski, V. V.  
894 Krzhizhanovskaya, J. J. Dongarra, and P. M. Slood, editors, *Computational Science – ICCS 2023*,  
895 pages 367–381, Cham, 2023c. Springer Nature Switzerland. ISBN 978-3-031-36027-5.
- 896 D. Bolin and J. Wallin. Spatially adaptive covariance tapering. *Spatial Statistics*, 18:163–178,  
897 2016. ISSN 2211-6753. doi: <https://doi.org/10.1016/j.spasta.2016.03.003>. URL <https://www.sciencedirect.com/science/article/pii/S2211675316000245>. Spatial Statistics  
898 Avignon: Emerging Patterns.
- 900 P. Boudier, A. Fillion, S. Gratton, S. Gürol, and S. Zhang. Data assimilation networks. *Journal*  
901 *of Advances in Modeling Earth Systems*, 15(4):e2022MS003353, 2023. doi: [https://doi.org/](https://doi.org/10.1029/2022MS003353)  
902 [10.1029/2022MS003353](https://doi.org/10.1029/2022MS003353). URL [https://agupubs.onlinelibrary.wiley.com/doi/abs/](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2022MS003353)  
903 [10.1029/2022MS003353](https://doi.org/10.1029/2022MS003353). e2022MS003353 2022MS003353.
- 904 J. Brajard, A. Carrassi, M. Bocquet, and L. Bertino. Combining data assimilation and  
905 machine learning to infer unresolved scale parametrization, 2021. URL [https://](https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2020.0086)  
906 [royalsocietypublishing.org/doi/abs/10.1098/rsta.2020.0086](https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2020.0086).
- 907 A. Carrassi, M. Bocquet, L. Bertino, and G. Evensen. Data assimilation in the geosciences: An  
908 overview of methods, issues, and perspectives. *WIREs Climate Change*, 9(5):e535, 2018. doi:  
909 <https://doi.org/10.1002/wcc.535>. URL [https://onlinelibrary.wiley.com/doi/abs/10.](https://onlinelibrary.wiley.com/doi/abs/10.1002/wcc.535)  
910 [1002/wcc.535](https://doi.org/10.1002/wcc.535).
- 911 Y. Chen, T. Davis, W. Hager, and S. Rajamanickam. Algorithm 887: Cholmod, supernodal sparse  
912 cholesky factorization and update/downdate. *ACM Trans. Math. Softw.*, 35, 01 2008.
- 913 S. Cheng and M. Qiu. Observation error covariance specification in dynamical systems for data  
914 assimilation using recurrent neural networks. *Neural Computing and Applications*, 34(16):  
915 13149–13167, Aug 2022. ISSN 1433-3058. doi: [10.1007/s00521-021-06739-4](https://doi.org/10.1007/s00521-021-06739-4). URL [https://](https://doi.org/10.1007/s00521-021-06739-4)  
916 [doi.org/10.1007/s00521-021-06739-4](https://doi.org/10.1007/s00521-021-06739-4).
- 917 S. Cheng, C. Quilodran-Casas, S. Ouala, A. Farchi, C. Liu, P. Tandeo, R. Fablet, D. Lucor, B. Iooss,  
918 J. Brajard, D. Xiao, T. Janjic, W. Ding, Y. Guo, A. Carrassi, M. Bocquet, and R. Arcucci. Machine

- 919 learning with data assimilation and uncertainty quantification for dynamical systems: a review,  
920 2023.
- 921 J. Chilès and P. Delfiner. *Geostatistics : modeling spatial uncertainty*. Wiley, New-York, second  
922 edition, 2012.
- 923 J.-P. Chilès and N. Desassis. *Fifty Years of Kriging*, pages 589–612. Springer International  
924 Publishing, Cham, 2018. ISBN 978-3-319-78999-6. doi: 10.1007/978-3-319-78999-6\_29.  
925 URL [https://doi.org/10.1007/978-3-319-78999-6\\_29](https://doi.org/10.1007/978-3-319-78999-6_29).
- 926 L. Clarotto, D. Allard, T. Romary, and N. Desassis. The spde approach for spatio-temporal datasets  
927 with advection and diffusion, 2022. URL <https://arxiv.org/abs/2208.14015>.
- 928 F. Counillon and L. Bertino. Ensemble optimal interpolation: multivariate properties in the gulf of  
929 mexico. *Tellus A*, 61(2):296–308, 2009. doi: <https://doi.org/10.1111/j.1600-0870.2008.00383.x>.  
930 x. URL [https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0870.2008.  
931 00383.x](https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0870.2008.00383.x).
- 932 N. Cressie and C. Wikle. *Statistics for Spatio-Temporal Data*. John Wiley & Sons, 2015. ISBN  
933 978-1-119-24304-5.
- 934 L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real nvp, 2017.
- 935 A. Doury, S. Somot, S. Gadat, A. Ribes, and L. Corre. Regional climate model emulator based  
936 on deep learning: concept and first evaluation of a novel hybrid downscaling approach. *Climate  
937 Dynamics*, 60(5):1751–1779, Mar 2023. ISSN 1432-0894. doi: 10.1007/s00382-022-06343-9.  
938 URL <https://doi.org/10.1007/s00382-022-06343-9>.
- 939 C. Dufau, M. Orszynowicz, G. Dibarboure, R. Morrow, and P.-Y. Le Traon. Mesoscale resolution  
940 capability of altimetry: Present and future. *Journal of Geophysical Research: Oceans*, 121(7):  
941 4910–4927, 2016. doi: 10.1002/2015JC010904. URL [https://agupubs.onlinelibrary.  
942 wiley.com/doi/abs/10.1002/2015JC010904](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015JC010904).
- 943 R. E. Ewing and H. Wang. A summary of numerical methods for time-dependent advection-  
944 dominated partial differential equations. *Journal of Computational and Applied Mathematics*,  
945 128(1):423–445, 2001. ISSN 0377-0427. doi: [https://doi.org/10.1016/S0377-0427\(00\)00522-7](https://doi.org/10.1016/S0377-0427(00)00522-7).

- 946 URL <https://www.sciencedirect.com/science/article/pii/S0377042700005227>.  
947 Numerical Analysis 2000. Vol. VII: Partial Differential Equations.
- 948 D. Esteban-Fernandez. Swot project mission performance and error budget document. Technical  
949 report, JPL, NASA, 2014.
- 950 G. Evensen. *Data Assimilation*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.  
951 ISBN 9783642037108 9783642037115. URL [http://link.springer.com/10.1007/](http://link.springer.com/10.1007/978-3-642-03711-5)  
952 [978-3-642-03711-5](http://link.springer.com/10.1007/978-3-642-03711-5).
- 953 G. Evensen, , F. C. Vossepoel, and P. J. van Leeuwen. *Data Assimilation Fundamentals*. Springer  
954 Textbooks in Earth Sciences, Geography and Environment (STEGE). Springer Cham, 2022.  
955 ISBN 978-3-030-96708-6. doi: 10.1007/978-3-030-96709-3. URL [https://doi.org/10.](https://doi.org/10.1007/978-3-030-96709-3)  
956 [1007/978-3-030-96709-3](https://doi.org/10.1007/978-3-030-96709-3).
- 957 R. Fablet, L. Drumetz, and F. Rousseau. Joint learning of variational representations and solvers  
958 for inverse problems with partially-observed data. *arXiv:2006.03653 [physics]*, 2020. URL  
959 <https://arxiv.org/abs/2006.03653>. arXiv:2006.03653.
- 960 R. Fablet, B. Chapron, L. Drumetz, E. Mémin, O. Pannekoucke, and F. Rousseau.  
961 Learning variational data assimilation models and solvers. *Journal of Advances in*  
962 *Modeling Earth Systems*, 13(10):e2021MS002572, 2021. doi: [https://doi.org/10.1029/](https://doi.org/10.1029/2021MS002572)  
963 [2021MS002572](https://doi.org/10.1029/2021MS002572). URL [https://agupubs.onlinelibrary.wiley.com/doi/abs/10.](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021MS002572)  
964 [1029/2021MS002572](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021MS002572). e2021MS002572 2021MS002572.
- 965 Q. Febvre, R. Fablet, J. L. Sommer, and C. Ubelmann. Joint calibration and mapping of satellite  
966 altimetry data using trainable variational models. In *ICASSP 2022 - 2022 IEEE International*  
967 *Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1536–1540, 2022. doi:  
968 [10.1109/ICASSP43922.2022.9746889](https://doi.org/10.1109/ICASSP43922.2022.9746889).
- 969 B. A. Finlayson. *Numerical Methods for Problems with Moving Fronts*. Ravenna Park Publishing,  
970 Seattle, 1992.
- 971 G.-A. Fuglstad, F. Lindgren, D. Simpson, and H. Rue. Exploring a new class of non-stationary  
972 spatial gaussian random fields with varying local anisotropy. *Statistica Sinica*, 25(1):115–133,  
973 2015a. ISSN 1017-0405. doi: 10.5705/ss.2013.106w.

- 974 G.-A. Fuglstad, D. Simpson, F. Lindgren, and H. Rue. Does non-stationary spatial data always  
975 require non-stationary random fields? *Spatial Statistics*, 14:505–531, 2015b. ISSN 2211-6753.  
976 doi: <https://doi.org/10.1016/j.spasta.2015.10.001>. URL [https://www.sciencedirect.com/  
977 science/article/pii/S2211675315000780](https://www.sciencedirect.com/science/article/pii/S2211675315000780).
- 978 G.-A. Fuglstad, D. Simpson, F. Lindgren, and H. Rue. Does non-stationary spatial data always  
979 require non-stationary random fields? *Spatial Statistics*, 14:505 – 531, 2015c. ISSN 2211-6753.  
980 doi: <https://doi.org/10.1016/j.spasta.2015.10.001>. URL [http://www.sciencedirect.com/  
981 science/article/pii/S2211675315000780](http://www.sciencedirect.com/science/article/pii/S2211675315000780).
- 982 R. Furrer, M. G. Genton, and D. Nychka. Covariance tapering for interpolation of large spatial  
983 datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523, 2006. ISSN  
984 10618600. URL <http://www.jstor.org/stable/27594195>.
- 985 L. Gaultier and C. Ubelmann. Swot simulator documentation. Technical report, JPL, NASA, 2010.
- 986 I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and  
987 Y. Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference  
988 on Neural Information Processing Systems - Volume 2, NIPS' 14*, page 2672–2680, Cambridge,  
989 MA, USA, 2014. MIT Press.
- 990 A. Grigorievskiy, N. D. Lawrence, and S. Särkkä. Parallelizable sparse inverse formulation gaussian  
991 processes (spingp). *2017 IEEE 27th International Workshop on Machine Learning for Signal  
992 Processing (MLSP)*, pages 1–6, 2016.
- 993 J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models, 2020.
- 994 Y. Huang, Y. Zhang, and J. A. Chambers. A novel kullback–leibler divergence minimization-based  
995 adaptive student’s t-filter. *IEEE Transactions on Signal Processing*, 67(20):5417–5432, 2019.  
996 doi: 10.1109/TSP.2019.2939079.
- 997 D. P. Kingma and M. Welling. Auto-encoding variational bayes, 2022.
- 998 K. Kurosawa and J. Poterjoy. A statistical hypothesis testing strategy for adaptively blending  
999 particle filters and ensemble kalman filters for data assimilation. *Monthly Weather Review*, 151

1000 (1):105 – 125, 2023. doi: 10.1175/MWR-D-22-0108.1. URL <https://journals.ametsoc.org/view/journals/mwre/151/1/MWR-D-22-0108.1.xml>.

1001

1002 F. Le Guillou. *Cartographie dynamique de la topographie de l’océan de surface par assimilation*  
1003 *de données altimétriques*. Theses, Université Grenoble Alpes [2020-....], May 2022. URL  
1004 <https://theses.hal.science/tel-03775828>.

1005 F. Le Guillou, L. Gaultier, M. Ballarotta, S. Metref, C. Ubelmann, E. Cosme, and M.-H. Rio.  
1006 Regional mapping of energetic short mesoscale ocean dynamics from altimetry: performances  
1007 from real observations. *EGUsphere*, 2023:1–17, 2023. doi: 10.5194/egusphere-2023-509. URL  
1008 <https://egusphere.copernicus.org/preprints/2023/egusphere-2023-509/>.

1009 H. Lewy, K. Friedrichs, and R. Courant. Über die partiellen differenzgleichungen der mathema-  
1010 tischen physik. *Mathematische Annalen*, 100:32–74, 1928. URL [http://eudml.org/doc/](http://eudml.org/doc/159283)  
1011 [159283](http://eudml.org/doc/159283).

1012 X. Li, B. Wu, X. Zhu, and H. Yang. Consecutively missing seismic data interpolation based on  
1013 coordinate attention unet. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. doi:  
1014 [10.1109/LGRS.2021.3128511](https://doi.org/10.1109/LGRS.2021.3128511).

1015 F. Lindgren, H. Rue, and J. Lindström. An explicit link between gaussian fields and gaussian  
1016 markov random fields: the stochastic partial differential equation approach. *Journal of the Royal*  
1017 *Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011. doi: 10.1111/j.  
1018 [1467-9868.2011.00777.x](https://doi.org/10.1111/j.1467-9868.2011.00777.x). URL [https://rss.onlinelibrary.wiley.com/doi/abs/10.](https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2011.00777.x)  
1019 [1111/j.1467-9868.2011.00777.x](https://doi.org/10.1111/j.1467-9868.2011.00777.x).

1020 S. A. Martin, G. E. Manucharyan, and P. Klein. Synthesizing sea surface temper-  
1021 ature and satellite altimetry observations using deep learning improves the accuracy  
1022 and resolution of gridded sea surface height anomalies. *Journal of Advances in*  
1023 *Modeling Earth Systems*, 15(5):e2022MS003589, 2023. doi: [https://doi.org/10.1029/](https://doi.org/10.1029/2022MS003589)  
1024 [2022MS003589](https://doi.org/10.1029/2022MS003589). URL [https://agupubs.onlinelibrary.wiley.com/doi/abs/10.](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2022MS003589)  
1025 [1029/2022MS003589](https://doi.org/10.1029/2022MS003589). e2022MS003589 2022MS003589.

1026 L. Menut, B. Bessagnet, R. Briant, A. Cholakian, F. Couvidat, S. Mailler, R. Pennel, G. Siour,  
1027 P. Tuccella, S. Turquety, and M. Valari. The chimere v2020r1 online chemistry-transport model.

- 1028 *Geoscientific Model Development*, 14(11):6781–6811, 2021. doi: 10.5194/gmd-14-6781-2021.  
1029 URL <https://gmd.copernicus.org/articles/14/6781/2021/>.
- 1030 S. Metref, E. Cosme, F. Le Guillou, J. Le Sommer, J.-M. Brankart, and J. Verron. Wide-  
1031 swath altimetric satellite data assimilation with correlated-error reduction. *Frontiers in*  
1032 *Marine Science*, 6:822, 2020. ISSN 2296-7745. doi: 10.3389/fmars.2019.00822. URL  
1033 <https://www.frontiersin.org/article/10.3389/fmars.2019.00822>.
- 1034 J.-M. Molines. meom-configurations/NATL60-CJM165: NATL60 code used for CJM165 experi-  
1035 ment, Mar. 2018. URL <https://zenodo.org/record/1210116/#.XDpmc89Khp8>.
- 1036 T. Nakamura, K. Fukami, K. Hasegawa, Y. Nabae, and K. Fukagata. Convolutional neural network  
1037 and long short-term memory based reduced order surrogate for minimal turbulent channel flow.  
1038 *Physics of Fluids*, 33(2), Feb. 2021. ISSN 1070-6631. doi: 10.1063/5.0039845. Publisher  
1039 Copyright: © 2021 Author(s).
- 1040 A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga,  
1041 and A. Lerer. Automatic differentiation in pytorch. 2017.
- 1042 M. Pereira, N. Desassis, and D. Allard. Geostatistics for large datasets on riemannian manifolds:  
1043 A matrix-free approach. *Journal of Data Science*, 20(4):512–532, 2022. ISSN 1680-743X. doi:  
1044 10.6339/22-JDS1075.
- 1045 H. S. Price, R. S. Varga, and J. E. Warren. Applications of oscillation matrices to diequations. *J.*  
1046 *Math. Phys*, 45, 1966.
- 1047 F. Rozet and G. Louppe. Score-based data assimilation. In A. Oh, T. Naumann,  
1048 A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural*  
1049 *Information Processing Systems*, volume 36, pages 40521–40541. Curran Associates,  
1050 Inc., 2023. URL [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/](https://proceedings.neurips.cc/paper_files/paper/2023/file/7f7fa581cc8a1970a4332920cdf87395-Paper-Conference.pdf)  
1051 [7f7fa581cc8a1970a4332920cdf87395-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/7f7fa581cc8a1970a4332920cdf87395-Paper-Conference.pdf).
- 1052 J. J. Ruiz, M. Pulido, and T. Miyoshi. Estimating model parameters with ensemble-based data  
1053 assimilation: A review. *Journal of the Meteorological Society of Japan. Ser. II*, 91(2):79–99,  
1054 2013.

- 1055 M. A. Sacco, J. J. Ruiz, M. Pulido, and P. Tandeo. Evaluation of machine learning techniques  
1056 for forecast uncertainty quantification. *Quarterly Journal of the Royal Meteorological Society*,  
1057 148(749):3470–3490, 2022. doi: <https://doi.org/10.1002/qj.4362>. URL [https://rmets.  
1058 onlinelibrary.wiley.com/doi/abs/10.1002/qj.4362](https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.4362).
- 1059 M. A. Sacco, M. Pulido, J. J. Ruiz, and P. Tandeo. On-line machine-learning forecast un-  
1060 certainty estimation for sequential data assimilation. *Quarterly Journal of the Royal Me-  
1061 teorological Society*, n/a(n/a), 2024. doi: <https://doi.org/10.1002/qj.4743>. URL [https:  
1062 //rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.4743](https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.4743).
- 1063 S. Särkka and J. Hartikainen. Infinite-dimensional kalman filtering approach to spatio-temporal  
1064 gaussian process regression. In N. D. Lawrence and M. Girolami, editors, *Proceedings of  
1065 the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of  
1066 *Proceedings of Machine Learning Research*, pages 993–1001, La Palma, Canary Islands, 21–23  
1067 Apr 2012. PMLR. URL <https://proceedings.mlr.press/v22/sarkka12.html>.
- 1068 S. Särkka, A. Solin, and J. Hartikainen. Spatiotemporal learning via infinite-dimensional bayesian  
1069 filtering and smoothing: A look at gaussian process regression through kalman filtering. *IEEE  
1070 Signal Processing Magazine*, 30(4):51–61, 2013. doi: 10.1109/MSP.2013.2246292.
- 1071 M. Seeger, A. Hetzel, Z. Dai, E. Meissner, and N. D. Lawrence. Auto-differentiating linear algebra,  
1072 2019.
- 1073 Z. Shi, X. Xu, X. Liu, J. Chen, and M.-H. Yang. Video frame interpolation transformer, 2022.  
1074 URL <https://arxiv.org/abs/2111.13817>.
- 1075 F. Sigrist, H. R. Künsch, and W. A. Stahel. Stochastic partial differential equation based modelling  
1076 of large space–time data sets. *Journal of the Royal Statistical Society: Series B (Statistical  
1077 Methodology)*, 77(1):3–33, 2015. doi: <https://doi.org/10.1111/rssb.12061>. URL [https://  
1078 rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12061](https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12061).
- 1079 H. Song, I. Hoteit, B. D. Cornuelle, X. Luo, and A. C. Subramanian. An adjoint-based adap-  
1080 tive ensemble kalman filter. *Monthly Weather Review*, 141(10):3343 – 3359, 2013. doi:  
1081 10.1175/MWR-D-12-00244.1. URL [https://journals.ametsoc.org/view/journals/  
1082 mwre/141/10/mwr-d-12-00244.1.xml](https://journals.ametsoc.org/view/journals/mwre/141/10/mwr-d-12-00244.1.xml).

- 1083 J. C. Strikwerda. *Finite Difference Schemes and Partial Differential Equations (Wadsworth &*  
1084 *Brooks/Cole Mathematics Series)*. Wadsworth & Brooks/Cole Advanced Books & Software,  
1085 Pacific Grove, 1989.
- 1086 Z. Su, J. Wang, P. Klein, A. F. Thompson, and D. Menemenlis. Ocean submesoscales as a  
1087 key component of the global heat budget. *Nature Communications*, 9(1):775, Feb 2018.  
1088 ISSN 2041-1723. doi: 10.1038/s41467-018-02983-w. URL [https://doi.org/10.1038/  
1089 s41467-018-02983-w](https://doi.org/10.1038/s41467-018-02983-w).
- 1090 G. Taburet, A. Sanchez-Roman, M. Ballarotta, M.-I. Pujol, J.-F. Legeais, F. Fournier, Y. Faugere,  
1091 and G. Dibarboure. DUACS DT2018: 25 years of reprocessed sea level altimetry products. 15  
1092 (5):1207–1224, 2019. ISSN 1812-0784. doi: <https://doi.org/10.5194/os-15-1207-2019>. URL  
1093 <https://www.ocean-sci.net/15/1207/2019/>. Publisher: Copernicus GmbH.
- 1094 P. Tandeo, P. Ailliot, J. Ruiz, A. Hannart, B. Chapron, A. Cuzol, V. Monbet, R. Easton, and  
1095 R. Fablet. Combining Analog Method and Ensemble Data Assimilation: Application to the  
1096 Lorenz-63 Chaotic System. In V. Lakshmanan, E. Gilleland, A. McGovern, and M. Tingley,  
1097 editors, *Machine Learning and Data Mining Approaches to Climate Science*, pages 3–12.  
1098 Springer, 2015.
- 1099 P. Tandeo, P. Ailliot, M. Bocquet, A. Carrassi, T. Miyoshi, M. Pulido, and Y. Zhen. A review of  
1100 innovation-based methods to jointly estimate model and observation error covariance matrices in  
1101 ensemble data assimilation. *Monthly Weather Review*, 148(10):3973–3994, 2020. doi: 10.1175/  
1102 MWR-D-19-0240.1. URL [https://journals.ametsoc.org/view/journals/mwre/148/  
1103 10/mwrD190240.xml](https://journals.ametsoc.org/view/journals/mwre/148/10/mwrD190240.xml).
- 1104 J. Tong, L. Xie, W. Yang, and K. Zhang. Probabilistic decomposition transformer for time series  
1105 forecasting, 2022.
- 1106 P.-Y. Traon, F. Nadal, and N. Ducet. An improved mapping method of mul-  
1107 tisatellite altimeter data. *Journal of Atmospheric and Oceanic Technology*, 15  
1108 (2):522–534, 1998. doi: 10.1175/1520-0426(1998)015<0522:AIMMOM>2.0.CO;  
1109 2. URL [https://journals.ametsoc.org/view/journals/atot/15/2/1520-0426\  
1110 \\_1998\\_015\\_0522\\_aimmom\\_2\\_0\\_co\\_2.xml](https://journals.ametsoc.org/view/journals/atot/15/2/1520-0426_1998_015_0522_aimmom_2_0_co_2.xml).



- 1111 C. Ubelmann, P. Klein, and L.-L. Fu. Dynamic Interpolation of Sea Surface Height and  
1112 Potential Applications for Future High-Resolution Altimetry Mapping. *Journal of Atmo-*  
1113 *spheric and Oceanic Technology*, 32(1):177–184, Oct. 2014. ISSN 0739-0572. doi: 10.  
1114 1175/JTECH-D-14-00152.1. URL [http://journals.ametsoc.org/doi/abs/10.1175/  
1115 JTECH-D-14-00152.1](http://journals.ametsoc.org/doi/abs/10.1175/JTECH-D-14-00152.1).
- 1116 C. Ubelmann, P. Klein, and L.-L. Fu. Dynamic interpolation of sea surface height and potential  
1117 applications for future high-resolution altimetry mapping. *Journal of Atmospheric and Oceanic*  
1118 *Technology*, 32(1):177 – 184, 2015. doi: 10.1175/JTECH-D-14-00152.1. URL [https://  
1119 journals.ametsoc.org/view/journals/atot/32/1/jtech-d-14-00152\\_1.xml](https://journals.ametsoc.org/view/journals/atot/32/1/jtech-d-14-00152_1.xml).
- 1120 C. Ubelmann, G. Dibarboure, L. Gaultier, A. Ponte, F. Ardhuin, M. Ballarotta, and Y. Faugère.  
1121 Reconstructing ocean surface current combining altimetry and future spaceborne doppler data.  
1122 *Journal of Geophysical Research: Oceans*, 126(3):e2020JC016560, 2021. doi: [https://doi.org/  
1123 10.1029/2020JC016560](https://doi.org/10.1029/2020JC016560). URL [https://agupubs.onlinelibrary.wiley.com/doi/abs/  
1124 10.1029/2020JC016560](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020JC016560). e2020JC016560 2020JC016560.
- 1125 P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski,  
1126 P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman,  
1127 N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, Í. Polat, Y. Feng,  
1128 E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero,  
1129 C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0  
1130 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature*  
1131 *Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.
- 1132 H. Wackernagel. *Multivariate Geostatistics. An Introduction with Applications*. Springer-Verlag  
1133 Berlin Heidelberg, New York, 2003. doi: 10.1007/978-3-662-05294-5. ISBN 978-3-540-  
1134 44142-7.
- 1135 P. Whittle. On stationary processes in the plane. *Biometrika*, 41(3-4):434–449, 1953. doi: 10.  
1136 1093/biomet/41.3-4.434. URL [http://biomet.oxfordjournals.org/content/41/3-4/  
1137 434.short](http://biomet.oxfordjournals.org/content/41/3-4/434.short).
- 1138 X. Zhang, M. A. Nawaz, X. Zhao, and A. Curtis. Chapter two - an introduction to variational  
1139 inference in geophysical inverse problems. In C. Schmelzbach, editor, *Inversion of Geophysical*

1140 *Data*, volume 62 of *Advances in Geophysics*, pages 73–140. Elsevier, 2021. doi: [https://](https://doi.org/10.1016/bs.agph.2021.06.003)  
1141 [doi.org/10.1016/bs.agph.2021.06.003](https://doi.org/10.1016/bs.agph.2021.06.003). URL [https://www.sciencedirect.com/science/](https://www.sciencedirect.com/science/article/pii/S0065268721000030)  
1142 [article/pii/S0065268721000030](https://www.sciencedirect.com/science/article/pii/S0065268721000030).

1143 Y. Zhen, P. Tandeo, S. Leroux, S. Metref, T. Penduff, and J. L. Sommer. An adaptive op-  
1144 timal interpolation based on analog forecasting: Application to ssh in the gulf of mex-  
1145 ico. *Journal of Atmospheric and Oceanic Technology*, 37(9):1697 – 1711, 2020. doi:  
1146 [10.1175/JTECH-D-20-0001.1](https://doi.org/10.1175/JTECH-D-20-0001.1). URL [https://journals.ametsoc.org/view/journals/](https://journals.ametsoc.org/view/journals/atot/37/9/jtechD200001.xml)  
1147 [atot/37/9/jtechD200001.xml](https://journals.ametsoc.org/view/journals/atot/37/9/jtechD200001.xml).