Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2017-46
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 7 August 2017

Nonlinear Processes in Geophysics
Discussions
Open Access

Optimal Transport for Variational Data Assimilation

Nelson Feyeux[1], Arthur Vidard[1], and Maëlle Nodet[1]

[1]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France

Correspondence to: A. Vidard (arthur.vidard@inria.fr)

Abstract. Variational data assimilation methods are designed to estimate an unknown initial condition of a model using observations. To do so, one needs to compare model outputs and observations. This is generally performed using Euclidean distances. This paper investigates another distance choice: the Wasserstein distance, stemming from optimal transport theory.

We develop a variational data assimilation method using this distance. We investigate the impact of the scalar product, the gradient choice as well as the minimization algorithm. With appropriate choices, we show successful results on preliminary experiments. Optimal-transport-based optimization seems to be promising to preserve the geometrical properties of the estimated initial condition.

1 Introduction

Understanding and forecasting the evolution of a given system is a crucial topic in an ever increasing number of application domains. To achieve that goal, one can rely on multiple sources of information, namely observations of the system, numerical model describing its behaviour, as well as additional *a priori* knowledge such as statistical information or previous forecasts. To combine these heterogeneous sources of observation it is common practice to use so-called data assimilation methods (e.g., see reference books Lewis et al. (2006); Law et al. (2015)). They aim at finding either the initial/boundary conditions or some parameters of a numerical model. They are extensively used in numerical weather forecasting for instance (e.g., see reviews in the books Park and Xu (2009, 2013)).

The estimation of the different elements to be sought (the control vector) is performed in data assimilation through the comparison between the observations and their model counterparts. The control vector should be adjusted such that its model outputs would fit the observations, while taking into account that these observations are unperfect and corrupted by noise and errors.

Data assimilation methods are divided into three disctinct classes. First, there is statistical filtering based on Kalman filters. Then, the variational data assimilation methods based on the optimal control theory. More recently an hybrid of both approaches have been developed (Hamill and Snyder, 2000; Buehner, 2005; Bocquet and Sakov, 2014). In this paper we focus on the variational data assimilation. It consists in minimizing a cost function written as the distance between the observations and their model counterparts. A Tikhonov regularization is also added and so the distance between the control vector and a background state carying the *a priori* information is added in the cost function.

Thus the cost function contains the misfit between the data (*a priori* and observations) and their control and model counterparts. Minimizing the cost function aims to reach a compromise in which these errors are smallest as possible. The errors can be decomposed into amplitude and position errors. Position errors mean that the structural elements are present in the data, but misplaced. Some methods have been proposed in order to deal with position errors (Hoffman and Grassotti, 1996; Ravela et al., 2007). These involve a preprocessing step which consists in displacing the different data so they fit better with each other. Then the data assimilation is performed accounting for those displaced data.

A distance has to be chosen in order to compare the different data and measure the misfits. Usually, an Euclidean distance is used, often weighted to take into account the statistical errors. But Euclidean distances have trouble capturing position errors. This is illustrated in Fig. 1. The second density can be seen as the first one with position error. The middle point between the two densities in the sense of the L^2 distance (that is, the mean) has not the desired shape nor the desired localization. We investigate in this article the idea of using instead a distance stemming from the optimal transport theory, the Wasserstein distance, which can take into account position errors. In Fig. 1 we show that the mean with respect to the Wasserstein distance is what we want it to be: same shape, same amplitude, located in-between. It conserves the shape of the data. This is what we want to achieve when dealing with position errors.

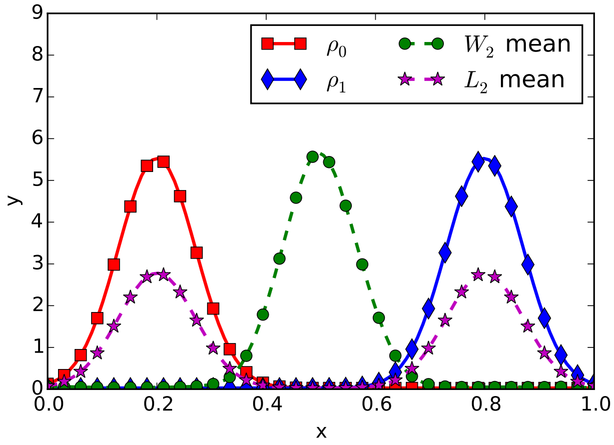

Figure 1. Wasserstein ($\mathcal{W}_2$) and Euclidean ($\mathcal{L}_2$) means of two densities ρ_0 and ρ_1.

The optimal transport theory has been founded by Monge in 1781 (Monge, 1781). He searched for the optimal way of displacing sand piles onto holes of the same volume, minimizing the total cost of displacement. This can be seen as a transportation problem between two probability densities. A modern presentation can be found in Villani (2003) and will be quickly recalled in Section 2.2.

Optimal transport has a wide spectrum of applications, from pure mathematical analysis to applied economics, from functional inequalities (Cordero-Erausquin et al., 2004) to the semi-geostrophic equations (Cullen and Gangbo, 2001), through astrophysics (Brenier et al., 2003), medicine (Ratner et al., 2015), crowd motion (Maury et al., 2010) or urban planning (Buttazzo and Santambrogio, 2005). From optimal transport theory several distances can be derived, the most widely known being the Wasserstein distance (denoted $\mathcal{W}_2$) which is sensitive to misplaced features, and is the primary focus of this paper. This distance is also widely used in computer vision, for example in classification of images (Rubner et al., 1998, 2000), interpolation Bonneel et al. (2011), or movie reconstruction (Delon and Desolneux, 2010). More recently, Farchi et al. (2016) used the Wasserstein distance is to compare observation and model simulations in an air pollution context, which is a first step toward data assimilation Actual use of optimal transport in a variational data assimilation has been proposed by Ning et al. (2014), to tackle model error. The authors use the Wasserstein distance instead of the classical L^2 norm *for model error control* in the cost function, and they offer promising results. Our contribution is in essence similar to them, in the fact that the Wasserstein distance is proposed in place of the L^2 distance. Looking more closely, we investigate a different question, namely the idea of using the Wasserstein distance to measure *the observation misfit*. Also, we underline and investigate the impact of the choice of the scalar products, gradient formulations, as well as minimization algorithm choices on the assimilation performance, which is not discussed in Ning et al. (2014). This particularly subtle mathematical consideration is indeed crucial for the algorithm convergence, as will be shown in this paper, and is our main contribution.

The goal of the paper is to perform variational data assimilation with a cost function written with the Wasserstein distance. It may be extended to other type of data assimilation methods such as filtering methods but it largely exceeds the scope of this paper.

The present paper is organized as follows: first, in Section 2, variational data assimilation as well as Wasserstein distance are defined, and the ingredients required for the sequel are presented. The core of our contribution lays in Section 3: we first present the Wasserstein cost function, then we propose two choices for its gradients, as well as two optimization strategies for the minimization. Section 4 numerical illustrations are presented, choices for the gradients and the optimization methods are compared. Also, some difficulties related to the use of optimal transport will be pointed out and solutions proposed.

2 Materials and Methodology

The section deals with the presentation of variational data assimilation materials on the one hand, and optimal transport and Wasserstein distance materials on the other hand. Section 3 will combine both worlds and will constitute the core of our original production.

2.1 Variational data assimilation

Let us assume that a system state is described by a variable $\mathbf{x}$. We are also given observations $\mathbf{y}^{\text{obs}}$ of the system, which might be indirect, uncomplete and approximate. The state and the observations are linked by an operator $\mathcal{G}$ mapping the system state $\mathbf{x}$ to the observation space, so that the mathematical nature of $\mathcal{G}(\mathbf{x})$ and $\mathbf{y}^{\text{obs}}$ are the same. Data assimilation aims to find a good estimate of $\mathbf{x}$ using the observations $\mathbf{y}^{\text{obs}}$ and the knowledge of the operator $\mathcal{G}$. Variational data assimilation methods do so by finding the minimizer $\mathbf{x}$ of the misfit function $\mathcal{J}$ (the cost function) between the observations $\mathbf{y}^{\text{obs}}$ and their computed counterparts $\mathcal{G}(\mathbf{x})$,

$$\mathcal{J}(\mathbf{x}) = d_R(\mathcal{G}(\mathbf{x}), \mathbf{y}^{\text{obs}})^2$$

with d_R some distance to be precised. Generally, this problem is ill-posed. For the minimizer of $\mathcal{J}$ to be unique, a background term is added and acts like a Tikhonov regularization. This background term is generally expressed as the distance with a background term $\mathbf{x}^b$ which contains *a priori* informations. The actual cost function then writes

$$\mathcal{J}(\mathbf{x}) = d_R(\mathcal{G}(\mathbf{x}), \mathbf{y}^{\text{obs}})^2 + d_B(\mathbf{x}, \mathbf{x}^b)^2, \tag{1}$$

with d_B another distance to be specified. The control of $\mathbf{x}$ is done by the minimization of $\mathcal{J}$. Such minimization is generally carried out numerically using gradient descent methods. Paragraph 3.3 will give more details about the minimization process.

The distances to the observations d_R and to the background term d_B have to be chosen in this formulation. Usually, Euclidean distances ($\mathcal{L}^2$ distances, potentially weighted) are chosen, giving the following Euclidean cost function

$$\mathcal{J}(\mathbf{x}) = \|\mathcal{G}(\mathbf{x}) - \mathbf{y}^{\text{obs}}\|_2^2 + \|\mathbf{x} - \mathbf{x}^b\|_2^2, \tag{2}$$

with $\|\cdot\|_2$ the $\mathcal{L}^2$ norm defined by

$$\|\mathbf{x}\|_2^2 := \int |\mathbf{x}(x)|^2 \,\mathrm{d}x. \tag{3}$$

The Wasserstein distance $\mathcal{W}_2$ in place of d_R and d_B in equation (1) is another choice and will be investigated in the following. Such a cost function will be presented in Section 3. The Wasserstein distance is presented and defined in the following subsection.

2.2 Optimal transport and Wasserstein distance

The essentials of optimal transport theory and Wasserstein distance required for data assimilation are presented.

We define, in this order, the space of probability densities where the Wasserstein distance is defined, then the Wasserstein distance and finally the Wasserstein scalar product, a key ingredient for variational assimilation.

2.2.1 Probability densities

We consider the case where the observations can be represented as densities. A density is a non-negative function of space. For example, a grey-scaled image is a density, it can be seen as a function of space to $[0,1]$ where 0 encodes black and 1 encodes white.

Definition 2.1. Let Ω be a closed, convex, bounded set of $\mathbb{R}^d$ and let define the set of probability densities $\mathcal{P}(\Omega)$ be the set of non-negative functions of total mass 1:

$$\mathcal{P}(\Omega) := \left\{ \rho \geq 0 \colon \int\limits_{\Omega} \rho(x)\,\mathrm{d}x = 1 \right\}. \tag{4}$$

2.2.2 Wasserstein distance

The optimal transport problem is to compute among all the transportations between two probability densities, the one minimizing the kinetic energy. A transportation between two probability densities ρ_0 and ρ_1 is given by a time path $\rho(t,x)$ such that $\rho(t=0) = \rho_0$ and $\rho(t=1) = \rho_1$, and a velocity field $\mathbf{v}(t,x)$ such that the continuity equation holds,

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho \mathbf{v}) = 0. \tag{5}$$

Such a path $\rho(t)$ can be seen as interpolating ρ_0 and ρ_1. For $\rho(t)$ to stay in $\mathcal{P}(\Omega)$, the velocity field $\mathbf{v}(t,x)$ has to be tangent to the domain boundary, meaning that $\rho(t,x)\mathbf{v}(t,x)\cdot \boldsymbol{n}(x) = 0$ for almost all $(t,x) \in [0,1]\times\partial\Omega$. With this condition, the support of $\rho(t)$ remains in Ω.

The Wasserstein distance $\mathcal{W}_2$ is hence the minimum in terms of kinetic energy among all the transportations between ρ_0 and ρ_1,

$$\mathcal{W}_2^2(\rho_0,\rho_1) = \min_{(\rho,\mathbf{v})\in C(\rho_0,\rho_1)} \iint\limits_{[0,1]\times\Omega} \rho(t,x)|\mathbf{v}(t,x)|^2\,\mathrm{d}t\mathrm{d}x \tag{6}$$

with $C(\rho_0,\rho_1)$ representing the set of continuous transportations between ρ_0 and ρ_1 described by a velocity field $\mathbf{v}$ tangent to the boundary of the domain,

$$C(\rho_0,\rho_1) := \left\{ (\rho,\mathbf{v}) \text{ s.t. } \begin{array}{l} \partial_t\rho + \operatorname{div}(\rho\mathbf{v}) = 0, \\ \rho(t=0) = \rho_0,\ \rho(t=1) = \rho_1, \\ \rho\mathbf{v}\cdot\boldsymbol{n} = 0 \text{ on } \partial\Omega \end{array} \right\}. \tag{7}$$

This definition of the Wasserstein distance is the Benamou-Brenier formulation Benamou and Brenier (2000). There exist other definitions, based on the transport map or the transference plans, but slightly out of the scope of this article. See the introduction of Villani (2003) for more details.

Remark 2.2 (Minimizer and Kantorovich potential)**.** A remarkable point is that the optimal velocity field $\mathbf{v}$ is of the form

$$\mathbf{v}(t,x) = \nabla\Phi(t,x)$$

with Φ following the Hamilton-Jacobi equation Ambrosio et al. (2008)

$$\partial_t\Phi + \frac{|\nabla\Phi|^2}{2} = 0. \tag{8}$$

The equation of the optimal ρ is the continuity equation using this velocity field. Moreover, the function $\Psi(x) := -\Phi(t=0,x)$ is said to be the **Kantorovich potential** of the transport between ρ_0 and ρ_1. It is a useful feature in the derivation of the Wasserstein cost function presented in Section 3.

Finally, a few words should be said about the numerical computation of the Wasserstein distance. In one dimension, it is easy to compute as it has an exact solution: the Kantorovich potential Ψ of the transport between ρ_0 and ρ_1 solves $F_1(x - \nabla\Psi(x)) = F_0(x)$ for all x, with F_i the cumulative distribution function of ρ_i. For problems in more dimensions, there exists no general formula and more complex algorithms have to be used, like the primal-dual Papadakis et al. (2014) or the semi-discrete Mérigot (2011).

2.2.3 Wasserstein inner product

The scalar product between two functions is required for data assimilation and optimization: as we will recall later, the scalar product choice conditions the gradient value. This paper will consider the classical $\mathcal{L}^2$ scalar product as well as the one associated to the Wasserstein distance. A scalar product defines the angle and norm of vectors tangent to $\mathcal{P}(\Omega)$ at a point ρ_0. First, a tangent vector in ρ_0 is the derivative of a curve $\rho(t)$ passing through ρ_0. As a curve $\rho(t)$ can be described by a continuity equation, the space of tangent vectors, the tangent space, shall formally be defined by (cf. Otto (2001)),

$$T_{\rho_0}\mathcal{P} = \left\{ -\mathrm{div}(\rho_0 \nabla\Phi) \in \mathcal{L}^2(\Omega), \rho_0 \frac{\partial \Phi}{\partial \boldsymbol{n}} = 0 \text{ on } \partial\Omega \right\}. \tag{9}$$

Let us first recall that the Euclidean, or $\mathcal{L}^2$, scalar product $\langle \cdot, \cdot \rangle_2$ is defined on $T_{\rho_0}\mathcal{P}$ by

$$\forall \eta, \eta' \in T_{\rho_0}\mathcal{P}(\Omega), \quad \langle \eta, \eta' \rangle_2 := \int\limits_\Omega \eta(x)\eta'(x)\,\mathrm{d}x. \tag{10}$$

While the Wasserstein inner product $\langle \cdot, \cdot \rangle_W$ is defined for $\eta = -\mathrm{div}(\rho_0 \nabla\Phi), \eta' = -\mathrm{div}(\rho_0 \nabla\Phi') \in T_{\rho_0}\mathcal{P}$ by

$$\langle \eta, \eta' \rangle_W := \int\limits_\Omega \rho_0 \nabla\Phi \cdot \nabla\Phi' \,\mathrm{d}x. \tag{11}$$

One has to note that the inner product is dependent on $\rho_0 \in \mathcal{P}(\Omega)$. Finally, the norm associated to a tangent vector $\eta = -\mathrm{div}(\rho_0 \nabla\Phi) \in T_{\rho_0}\mathcal{P}$ is

$$\|\eta\|_W^2 = \int\limits_\Omega \rho_0 |\nabla\Phi|^2 \,\mathrm{d}x \tag{12}$$

hence the kinetic energy of the small displacement η. This point makes the link between this inner product and the Wasserstein distance.

3 Optimal transport-based data assimilation

This section is our main contribution. First we will consider the Wasserstein distance to compute the *observation term of the cost function*; second we will investigate the role of the scalar product choice as well as the gradient descent method on the assimilation algorithm efficiency.

3.1 Wasserstein cost function

In the framework of Section 2.2 we will define the data assimilation cost function using the Wasserstein distance. For this cost function to be well defined we assume that the control variables belong to $\mathcal{P}(\Omega)$ and that the observation variables belong to another space $\mathcal{P}(\Omega_o)$ with Ω_o a closed, convex, bounded set of $\mathbb{R}^{d'}$. Let us recall that this means that they are all non-negative densities of integral equal to 1. Having elements of integral 1 (or constant integral) may seem restrictive. Removing it is yet possible by using a modified version of the Wasserstein distance, presented for example in Chizat et al. (2015). For simplicity we do not consider here this possible generalization and all data have the same integral. The cost function (1) is rewritten using the Wasserstein distance defined in Section 2.2,

$$\mathcal{J}_{\mathcal{W}}(\mathbf{x}_0) = \frac{1}{2}\sum_{i=1}^{N^{\text{obs}}} \mathcal{W}_2^2(\mathcal{G}_i(\mathbf{x}_0), \mathbf{y}_i^{\text{obs}}) + \frac{\omega_b}{2}\mathcal{W}_2^2(\mathbf{x}_0, \mathbf{x}_0^b), \tag{13}$$

with $\mathcal{G}_i\colon \mathcal{P}(\Omega) \to \mathcal{P}(\Omega_o)$ the observation operator computing the $\mathbf{y}_i^{\text{obs}}$ counterpart from $\mathbf{x}_0$.

The variables $\mathbf{x}_0$ and $\mathbf{y}_i^{obs}$ may be vectors whose components are functions belonging repectively to $\mathcal{P}(\Omega)$ and $\mathcal{P}(\Omega_o)$. The Wasserstein distance between two such vectors is the sum of the distances between their components. The remainder of the article is easily adaptable to this case, but for simplicity we set $\mathbf{x}_0 = \rho_0 \in \mathcal{P}(\Omega)$ and $\mathbf{y}_i^{obs} = \rho_i^{\text{obs}} \in \mathcal{P}(\Omega)$. The Wasserstein cost function (13) then becomes

$$\mathcal{J}_{\mathcal{W}}(\rho_0) = \frac{1}{2}\sum_{i=1}^{N^{\text{obs}}} \mathcal{W}_2^2(\mathcal{G}_i(\rho_0), \rho_i^{\text{obs}}) \quad + \quad \frac{\omega_b}{2}\,\mathcal{W}_2^2(\rho_0, \rho_0^b). \tag{14}$$

To find the minimum of $\mathcal{J}_{\mathcal{W}}$, a gradient descent method is applied. It is presented in Section 3.3. As this type of algorithms requires the gradient of the cost function, the computation of the gradient of $\mathcal{J}_{\mathcal{W}}$ is the focus of the next Section.

3.2 Gradient of $\mathcal{J}_{\mathcal{W}}$

If $\mathcal{J}_{\mathcal{W}}$ is differentiable, its gradient is not unique but depends on the choice of a scalar product $\langle\cdot,\cdot\rangle$. Indeed, a gradient g of $\mathcal{J}_{\mathcal{W}}$ is such that

$$\forall \eta \in T_{\rho_0}\mathcal{P}, \quad \lim_{\epsilon \to 0} \frac{\mathcal{J}_{\mathcal{W}}(\rho_0 + \epsilon\eta) - \mathcal{J}_{\mathcal{W}}(\rho_0)}{\epsilon} = \langle \eta, g\rangle \tag{15}$$

An important consequence of this formula is: *choosing another scalar product will give another gradient*. Moreover, the choice of the scalar product and gradient is important as it can affect significantly the behavior of gradient descent algorithms, as will be illustrated in the numerical results in Section 4.

The classical $\mathcal{L}^2$ inner product will be used, as well as the $\mathcal{W}_2$ one. The latter appears naturally as we deal with the Wasserstein distance (see definition in Section 2.2.3). The associated gradients are respectively denoted $\text{grad}_2\mathcal{J}_{\mathcal{W}}(\rho_0)$ and

$\text{grad}_W \mathcal{J}_\mathcal{W}(\rho_0)$ and are the only elements of the tangent space $T_{\rho_0}\mathcal{P}$ of $\rho_0 \in \mathcal{P}(\Omega)$ such that

$$\begin{aligned}\forall \eta \in T_{\rho_0}\mathcal{P}, \quad \lim_{\epsilon \to 0} \frac{\mathcal{J}_\mathcal{W}(\rho_0 + \epsilon\eta) - \mathcal{J}_\mathcal{W}(\rho_0)}{\epsilon} &= \langle \text{grad}_2 \mathcal{J}_\mathcal{W}(\rho_0), \eta \rangle_2 \\ &= \langle \text{grad}_W \mathcal{J}_\mathcal{W}(\rho_0), \eta \rangle_W. \end{aligned} \tag{16}$$

Here in the notations, the word "grad" is used for the gradient of a function while the spatial gradient is denoted by the nabla sign ∇. The gradients of $\mathcal{J}_\mathcal{W}$ are elements of $T_{\rho_0}\mathcal{P}$ and hence functions of space.

The following theorem allows to compute both gradients of $\mathcal{J}_\mathcal{W}$:

Theorem 3.1. *For $i \in \{1, \dots, N^{obs}\}$, let Ψ^i be the Kantorovich potential (see Remark 2.2) of the transport between $\mathcal{G}_i(\rho_0)$ and ρ_i^{obs}. Let Ψ^b be the Kantorovich potential of the transport map between ρ_0 and ρ_0^b. Then,*

$$\text{grad}_2 \mathcal{J}_\mathcal{W}(\rho_0) = \omega_b \Psi^b + \sum_{i=1}^{N^{obs}} \mathbf{G}_i^*(\rho_0).\Psi^i + c \tag{17}$$

with c such that the integral of $\text{grad}_2 \mathcal{J}_\mathcal{W}(\rho_0)$ is zero, and $\mathbf{G}_i^*$ the adjoint of $\mathcal{G}_i$ w.r.t. the $\mathcal{L}_2$ inner product (see definition reminder below). Assuming that $\text{grad}_2 \mathcal{J}_\mathcal{W}(\rho_0)$ has the no-flux boundary condition (see comment about this assumption below)

$$\rho_0 \frac{\partial \text{grad}_2 \mathcal{J}_\mathcal{W}(\rho_0)}{\partial \boldsymbol{n}} = 0 \textit{ on } \partial\Omega$$

then the gradient w.r.t. the Wasserstein inner product is

$$\text{grad}_W \mathcal{J}_\mathcal{W}(\rho_0) = -\text{div}\Big(\rho_0 \nabla [\text{grad}_2 \mathcal{J}_\mathcal{W}(\rho_0)]\Big). \tag{18}$$

A proof of this Theorem can be found in Appendix A.

Remark 3.2 (Adjoint reminder)**.** The adjoint $\mathbf{G}_i^*(\rho_0)$ is defined by the classical equality

$$\forall \eta, \mu \in T_{\rho_0}\mathcal{P}, \langle \mathbf{G}_i^*(\rho_0).\mu, \eta \rangle_2 = \langle \mu, \mathbf{G}_i(\rho_0).\eta \rangle_2 \tag{19}$$

where $\mathbf{G}_i[\rho_0]$ is the tangent model, defined by

$$\forall \eta \in T_{\rho_0}\mathcal{P}, \mathbf{G}_i(\rho_0).\eta := \lim_{\epsilon \to 0} \frac{\mathcal{G}_i(\rho_0 + \epsilon\eta) - \mathcal{G}_i(\rho_0)}{\epsilon}. \tag{20}$$

Remark 3.3 (Assumption of no-flux boundary condition)**.** The condition of no-flux at the boundary for $\text{grad}_2 \mathcal{J}_\mathcal{W}(\rho_0)$, that is

$$\rho_0 \frac{\partial \text{grad}_2 \mathcal{J}_\mathcal{W}(\rho_0)}{\partial \boldsymbol{n}} \text{ on } \partial\Omega$$

is not necessarily satisfied. The Kantorovich potentials respect this condition. Indeed, their spatial gradients are velocities thus thus tangent to the boundary, see the end of Section 2.2. But it may not be conserved through the mapping with the adjoint model, $\mathbf{G}_i^*(\rho_0)$. In the case where $\mathbf{G}_i^*(\rho_0)$ does not preserve this condition, the Wasserstein gradient is not of integral zero. A possible workaround is to use a product coming from the unbalanced Wasserstein distance of Chizat et al. (2015).

3.3 Minimization of $\mathcal{J}_{\mathcal{W}}$

The minimizer of $\mathcal{J}_{\mathcal{W}}$ defined in (14) is expected to be a good trade-off between both the observations and the background with respect to the Wasserstein distance and to have good properties, as shown in Fig. 1. It can be computed through an iterative gradient-based descent method. Such methods start from a control state ρ_0^0 and step-by-step update it using an iteration of the form

$$\rho_0^{n+1} = \rho_0^n - \alpha^n d^n \tag{21}$$

where α^n is a real number (the step) and d^n is a function (the descent direction), chosen such that $\mathcal{J}_{\mathcal{W}}(\rho_0^{n+1}) < \mathcal{J}_{\mathcal{W}}(\rho_0^n)$. In gradient-based descent methods, d^n can be equal to the gradient of $\mathcal{J}_{\mathcal{W}}$ (steepest descent method), or to a function of the gradient and d^{n-1} (conjugate gradient, quasi-Newton methods, ...). Under sufficient conditions on (α^n), the sequence (ρ_0^n) converges to a local minimizer. See Nocedal and Wright (2006) for more details.

Remark 3.4 (Note on descent with the Wasserstein gradient)**.** With the Wasserstein gradient (18), the descent of $\mathcal{J}_{\mathcal{W}}$ follows an iteration scheme of the form

$$\rho_0^{n+1} = \rho_0^n + \alpha^n \operatorname{div}(\rho_0^n \nabla \Phi^n). \tag{22}$$

A more transport-like iteration could be used instead,

$$\rho_0^{n+1} = (I - \alpha^n \nabla \Phi^n) \# \rho_0^n \tag{23}$$

with $\#$ the notation of the push-forward by a transport map: if $T\colon \Omega \to \Omega$ is a diffeomorphism, $\rho_1 := T\#\rho_0$ is defined as

$$\rho_1(T)\,|\det(\nabla T)| = \rho_0.$$

Iteration (23) is much more interesting as Fig. 2 shows, first because ρ_0^{n+1} stays non-negative whatever α^n, then because it allows the supports of ρ_0^n and ρ_0^{n+1} to be different. It is the one we will use after.

Iteration (23) is equivalent to (22) when α^n tends to 0. Indeed, it can be shown that $\rho(t) := (I - t\nabla\Phi_0)\#\rho_0$ is the equation of a geodesic and is solution of the system of equations

$$\begin{cases} \partial_t \rho + \operatorname{div}(\rho \nabla \Phi) = 0 \\ \partial_t \Phi + \dfrac{|\nabla \Phi|^2}{2} = 0. \end{cases} \tag{24}$$

with initial conditions $\rho(0,x) = \rho_0(x)$ and $\Phi(0,x) = \Phi_0(x)$, see (Villani, 2003, (5.61)). Therefore, (22) is just an explicit time-discretization of (24) and is equivalent to (23) for α^n tending to 0.

4 Numerical illustrations

Let us recall that in the data assimilation vocabulary, the word "analysis" refers to the minimizer of the cost function at the end of the data assimilation process.

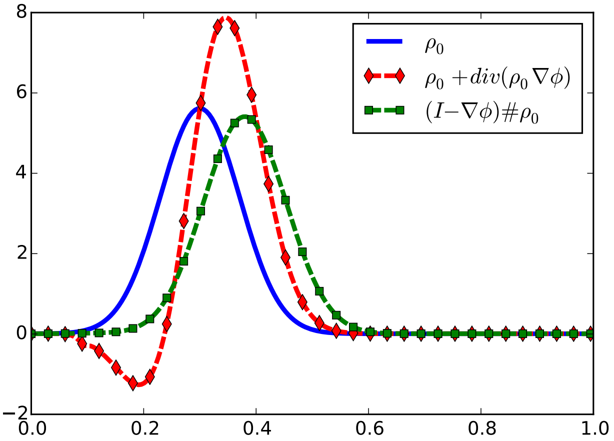

Figure 2. Comparison of iterations (22) and (23) with ρ_0 of limited support and Φ such that $\nabla\Phi$ is constant on the support of ρ_0.

In this section are presented the analyses resulting from the minimization of the Wasserstein cost function defined previously in (13), in particular when position errors occur. Results are compared with the results given by the $\mathcal{L}^2$ cost function defined in (2).

The experiments are all one-dimensional and $\Omega = [0,1]$. A first, simple experiment uses a linear observation operator $\mathcal{G}$. In a second experiment, the observation operator is non-linear, but results are still satisfactory.

Only a single variable is controlled. This variable ρ_0 represents the initial condition of an evolution problem. It is an element of $\mathcal{P}(\Omega)$, and observations are also elements of $\mathcal{P}(\Omega)$.

In this paper we chose to work in the twin experiments framework. In this context the true state, denoted ρ_0^t, is known and used to generate the observations: $\rho_i^{\text{obs}} = \mathcal{G}_i(\rho_0^t)$ at various times $(t_i)_{i=1..N_{\text{obs}}}$. The observations are perfect, that is noise-free and available everywhere in space. The background term is considered to have position errors only, and no amplitude error. Then, the data assimilation process aims to recover a good estimation of the true state, using the cost function involving the simulated observations and the background term. The analysis obtained after convergence can then be compared to the true state and effectiveness diagnostics can be made.

Both the Wasserstein and $\mathcal{L}^2$ cost functions are minimized through a steepest gradient method. The $\mathcal{L}^2$ gradient is used to minimize the $\mathcal{L}^2$ cost function. Both the $\mathcal{L}^2$ and $\mathcal{W}_2$ gradients are used for the Wasserstein cost functions, giving respectively, with $\Phi^n := \text{grad}_2 \mathcal{J}_{\mathcal{W}}(\rho_0^n)$, the iterations

$$\rho_0^{n+1} = \rho_0^n - \alpha^n \Phi^n \tag{DG2}$$

$$\rho_0^{n+1} = (I - \alpha^n \nabla\Phi^n)\#\rho_0^n. \tag{DG\#}$$

The value of α^n is chosen as optimal on each iteration and the algorithm stops when the decrement of $\mathcal{J}$ between two iterations is lower than 10^{-6}.

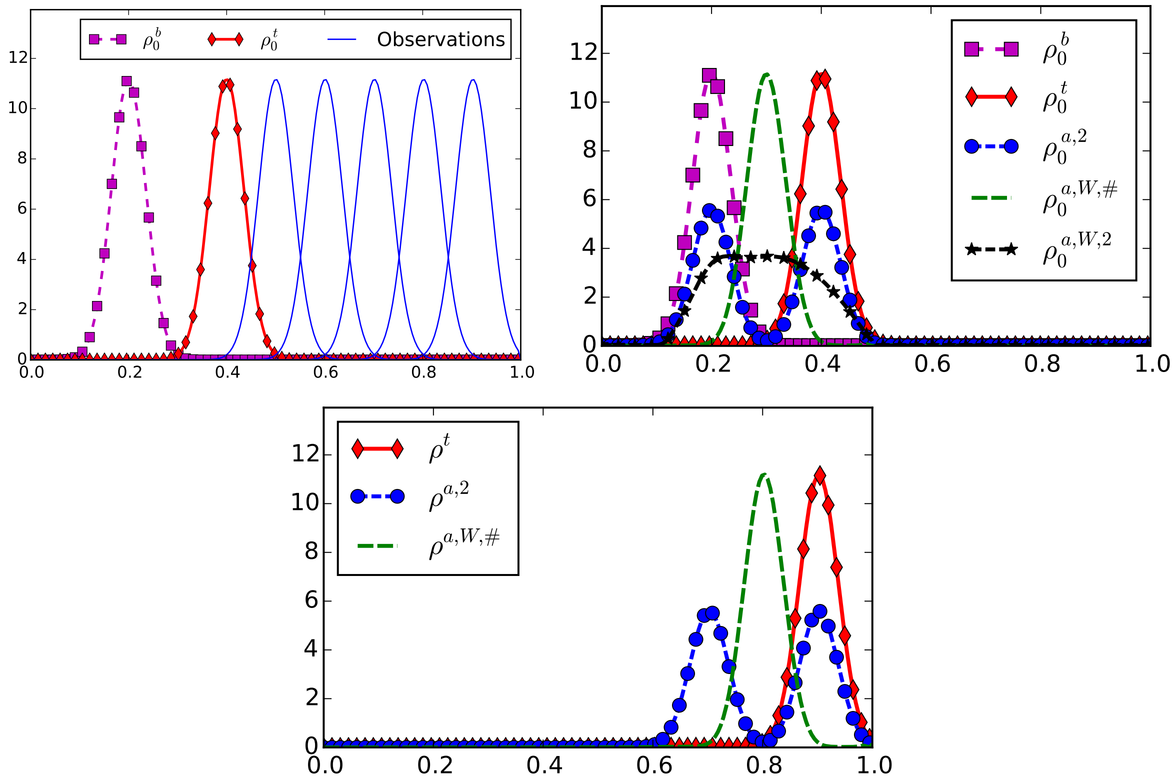

Figure 3. Top left: the twin experiments ingredients are plotted, namely true initial condition ρ_0^t, background term ρ_0^b, and observations at different times. Top right: we plot the analyses obtained after each proposed method, compared to ρ_0^b and ρ_0^t: $\rho_0^{a,2}$ corresponds to $\mathcal{J}_2$, $\rho_0^{a,W,2}$ to (DG2) and $\rho_0^{a,W,\#}$ to (DG#). Below are shown the outputs of the model, ρ^t, $\rho^{a,2}$ and $\rho^{a,W,\#}$, when taking respectively ρ_0^t, $\rho_0^{a,2}$ and $\rho_0^{a,W,\#}$ as initial condition.

4.1 Linear example

The first example involves a linear, evolution model as observation operators $(\mathcal{G}_i)_{i=1..N_{\text{obs}}}$ with the number of observations N_{obs} equal here to 5. Every single operator $\mathcal{G}_i$ maps an initial condition ρ_0 to $\rho(t_i)$ according to the following continuity equation

$$\partial_t \rho + \nabla \rho = 0. \tag{26}$$

The operator $\mathcal{G}_i$ is linear. We control ρ_0 only. The true state $\rho_0^t \in \mathcal{P}(\Omega)$ is a Gaussian, while the background term is also a Gaussian but located at a different place, as if it had position errors. On Fig. 3 (top left) are plotted the true and background states as well as the observations at various times. The computed analysis $\rho_0^{a,2}$ for the $\mathcal{L}_2$ cost function is shown on Fig. 3 (top right). This Figure shows also the analyses $\rho_0^{a,W,2}$ and $\rho_0^{a,W,\#}$ corresponding respectively to the (DG2) and (DG#) algorithms minimizing the Wasserstein $\mathcal{J}_{\mathcal{W}}$ cost function.

The analyses $\rho_0^{a,W,2}$ and $\rho_0^{a,W,\#}$ are different even if they arise from the same cost function $\mathcal{J}_\mathcal{W}$, which highlights the need for a well-suited scalar-product.

As expected in the introduction, see e.g. Fig. 1, minimizing $\mathcal{J}_2$ leads to an analysis being the $\mathcal{L}^2$-average of the background and true states (hence two small gaussians), while $\mathcal{J}_\mathcal{W}$ leads to a satisfactorily shaped analysis in-between the background and true states.

Remark 4.1. The analysis $\rho_0^{a,W,\#}$ is actually close to the average of ρ_0^b and ρ_0^t in the sense of the Wasserstein distance, that is to say close to the middle point on the Wasserstein geodesic between ρ_0^b and ρ_0^t (see also Figure 1 for a representation of the exact average).

The issue of amplitude of the analysis of $\rho_0^{a,2}$ and the issue of position of $\rho_0^{a,W,\#}$ are not corrected by the time evolution of the model, as shows Fig. 3 (bottom). At the end of the assimilation window, each of both of the analyses still have discrepancies with the observations.

As a conclusion of this first test case, we managed to write and minimize a cost function which gives a relevant analysis, contrary to what we obtain with the classical Euclidean cost function, in case of position errors. We also noticed that the success of the minimization of $\mathcal{J}_\mathcal{W}$ was clearly dependant on the scalar product choice.

4.2 Non-linear example

Further results are shown when a non-linear model is used in place of $\mathcal{G}$. The procedure is the same as the first test case. The non-linear model used is the Shallow-Water system described by

$$\begin{cases} \partial_t h + \partial_x(hu) = 0 \\ \partial_t u + u\partial_x u + g\partial_x h = 0 \end{cases}$$

subject to initial conditions $h(0) = h_0$ and $u(0) = u_0$, with reflective boundary conditions ($u|_{\partial\Omega} = 0$), where the constant g is the gravity acceleration. The variable h represents the water surface elevation, and u is the current velocity. If h_0 belongs to $\mathcal{P}(\Omega)$, then the corresponding solution $h(t)$ belongs to $\mathcal{P}(\Omega)$.

The true state is (h_0^t, u_0^t), where u_0^t is equal to 0 and h_0^t is a given Gaussian. The initial velocity field is supposed to be known and therefore not included in the control vector. Only h_0 is controlled, using $N_{\text{obs}} = 5$ direct observations of h and a background term h_0^b, also a localized Gaussian like h_0^t.

Data assimilation is performed by minimizing either the $\mathcal{J}_2$ or the $\mathcal{J}_W$ cost functions described above. Thanks to the wisdom gained during the first experiment, the (DG#) algorithm only is used for the minimization of $\mathcal{J}_\mathcal{W}$.

In Fig. 4 (left) we present h_0^t, h_0^b as well as the analyses corresponding to $\mathcal{J}_2$ and $\mathcal{J}_W$: $h_0^{a,2}$ and $h_0^{a,W,\#}$. Analysis $h_0^{a,2}$ is close to the L^2-average of the true and background states, even at time $t > 0$, while $h_0^{a,W,\#}$ lies close to the Wasserstein geodesic between the background and true states, and hence has the same shape as them (see Remark 4.1).

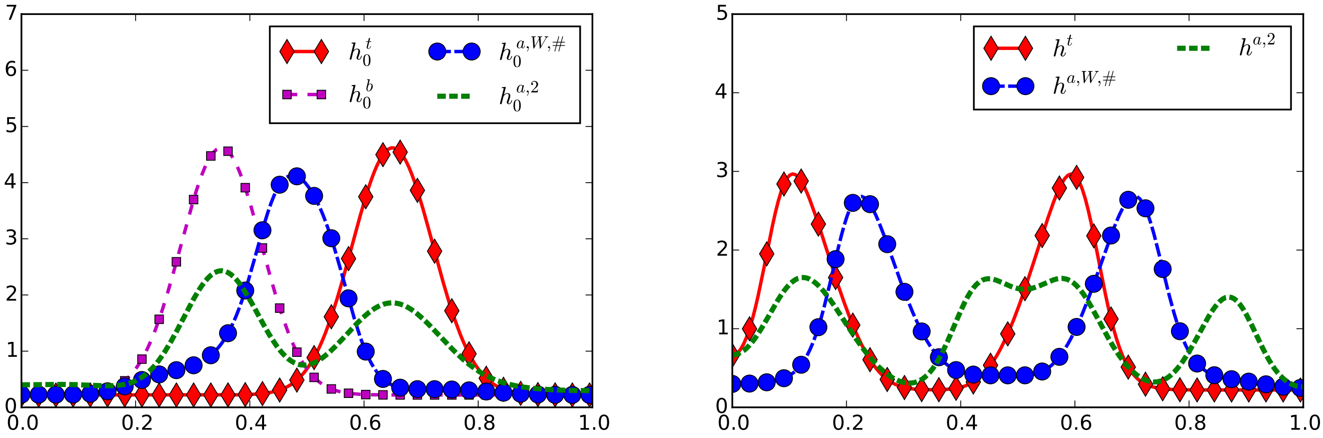

Figure 4. On the left are shown the true and background initial conditions, and the analyses $h_0^{a,2}$ and $h_0^{a,W}$ corresponding respectively to the Wasserstein and Euclidean cost functions to minimize. On the right we show the same plots (except the background one) but at the output of the model.

The Fig. 4 (right) shows that at the end of the assimilation window, the water height $h^{a,W,\#} = \mathcal{G}(h_0^{a,W,\#})$ is still more realistic than $h^{a,2} = \mathcal{G}(h_0^{a,2})$, when compared to the true state $h^t = \mathcal{G}(h_0^t)$.

The conclusion of this second test case is that even with non-linear models, our Wasserstein-based algorithm can give interesting results in case of position errors.

5 Conclusions

We showed through some examples that, if not taken into account, position errors can lead to unrealistic initial conditions when using classical variational data assimilation methods. Indeed, such methods use the Euclidean distance which can behave badly under position errors. To tackle this issue, we proposed instead the use of the Wasserstein distance to define the related cost function. The associated minimization algorithm was discussed and we showed that using descent iterations following Wasserstein geodesics lead to more consistent results.

On academic examples the corresponding cost function produces an analysis lying close to the Wasserstein average between the true and background states, and therefore has the same shape as them, and is well fit to correct position errors. This also gives more realistic predictions. This is a preliminary study, some issues have yet to be addressed for realistic applications, such as relaxing the constant-mass and positivity hypotheses and extending the problem to 2D applications.

Appendix A: Proof of Theorem 3.1

To prove Theorem 3.1, one first needs to differentiate the Wasserstein distance. The following Lemma from (Villani, 2003, Theorem 8.13 p.264) gives the gradient of the Wasserstein distance.

Lemma A.1 (Differentiation of the Wasserstein distance)**.** *Let $\rho_0, \rho_1 \in \mathcal{P}(\Omega)$, $\eta \in T_{\rho_0}\mathcal{P}$. For small enough $\epsilon \in \mathbb{R}$,*

$$\frac{1}{2}\mathcal{W}_2^2(\rho_0 + \epsilon\eta, \rho_1) = \frac{1}{2}\mathcal{W}_2^2(\rho_0, \rho_1) + \epsilon\langle \eta, \phi\rangle_2 + o(\epsilon) \tag{A1}$$

with $\phi(x)$ the Kantorovich potential of the transport between ρ_0 and ρ_1.

Proof of Theorem 3.1. Let $\rho_0 \in \mathcal{P}(\Omega)$ and $\eta = -\mathrm{div}(\rho_0 \nabla\Phi) \in T_{\rho_0}\mathcal{P}$. From the definition of $\mathcal{J}_{\mathcal{W}}$ in (13), from the defintion of the tangent model (20) and in application of the Lemma A.1,

$$\begin{aligned}\lim_{\epsilon\to 0}\frac{\mathcal{J}_{\mathcal{W}}(\rho_0+\epsilon\eta) - \mathcal{J}_{\mathcal{W}}(\rho_0)}{\epsilon} &= \sum_{i=1}^{N^{obs}} \langle \mathbf{G}_i[\rho_0]\eta, \phi^i\rangle_2 + \omega_b\langle \eta, \phi^b\rangle_2 \\ &= \left\langle \eta, \sum_{i=1}^{N^{obs}} \mathbf{G}_i^*[\rho_0]\phi^i + \omega_b\phi^b \right\rangle_2 \\ &= \left\langle \eta, \sum_{i=1}^{N^{obs}} \mathbf{G}_i^*[\rho_0]\phi^i + \omega_b\phi^b + c \right\rangle_2 \end{aligned} \tag{A2}$$

with c such that the integral of the right hand side term is zero, so that the right hand side term belongs to $T_{\rho_0}\mathcal{P}$. The $\mathcal{L}^2$ gradient of $\mathcal{J}_{\mathcal{W}}$ is thus

$$\mathrm{grad}_2\mathcal{J}_{\mathcal{W}}(\rho_0) = \sum_{i=1}^{N^{obs}} \mathbf{G}_i^*[\rho_0]\phi^i + \omega_b\phi^b + c \tag{A3}$$

To get the Wasserstein gradient of $\mathcal{J}_{\mathcal{W}}$, the same has to be done with the Wasserstein product. We let $\eta = -\mathrm{div}(\rho\nabla\Phi)$ and $g = \mathrm{grad}_2\mathcal{J}_{\mathcal{W}}(\rho_0)$ so that equations (A2) and (A3) give

$$\begin{aligned}\langle \eta, g\rangle_2 &= \langle -\mathrm{div}(\rho_0\nabla\Phi), g\rangle_2 \\ &= -\int_\Omega \mathrm{div}(\rho_0\nabla\Phi)g \\ &= \int_\Omega \rho_0\nabla\Phi\nabla g. \end{aligned} \tag{A4}$$

Last equality comes from Stokes theorem and from the fact that Φ is of zero normal derivative at the boundary. The last term gives the Wasserstein gradient because if g is with Neumann boundary conditions, we have

$$\int_\Omega \rho_0\nabla\Phi\nabla g = \langle \eta, -\mathrm{div}(\rho_0\nabla g)\rangle_W, \tag{A5}$$

hence

$$\forall \eta \in T_{\rho_0}\mathcal{P}, \quad \lim_{\epsilon \longrightarrow 0} \frac{\mathcal{J}_{\mathcal{W}}(\rho_0 + \epsilon\eta) - \mathcal{J}_{\mathcal{W}}(\rho_0)}{\epsilon} = \langle \eta, -\mathrm{div}(\rho_0 \nabla g) \rangle_W. \tag{A6}$$

□

Competing interests. No competing interest are present.

5 *Acknowledgements.* Nelson Feyeux is supported by the Région Rhône Alpes Auvergne through the ARC3 *Environment* PhD fellowship program.

References

Ambrosio, L., Gigli, N., and Savaré, G.: Gradient flows: in metric spaces and in the space of probability measures, Springer Science & Business Media, 2008.

Benamou, J.-D. and Brenier, Y.: A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem, Numerische Mathematik, 84, 375–393, 2000.

Bocquet, M. and Sakov, P.: An iterative ensemble Kalman smoother, Quaterly Journal of the Royal Meteorological Society, 140, 1521–1535, doi:10.1002/qj.2236, 2014.

Bonneel, N., Van De Panne, M., Paris, S., and Heidrich, W.: Displacement interpolation using Lagrangian mass transport, in: ACM Transactions on Graphics (TOG), vol. 30, p. 158, ACM, 2011.

Brenier, Y., Frisch, U., Hénon, M., Loeper, G., Matarrese, S., Mohayaee, R., and Sobolevskiĭ, A.: Reconstruction of the early Universe as a convex optimization problem, Monthly Notices of the Royal Astronomical Society, 346, 501–524, 2003.

Buehner, M.: Ensemble-derived stationary and flow-dependent background-error covariances: Evaluation in a quasi-operational NWP setting, Quaterly Journal of the Royal Meteorological Society, 131, 1013–1043, doi:doi: 10.1256/qj.04.15, 2005.

Buttazzo, G. and Santambrogio, F.: A model for the optimal planning of an urban area, SIAM Journal on Mathematical Analysis, 37, 514–530, 2005.

Chizat, L., Schmitzer, B., Peyré, G., and Vialard, F.-X.: An Interpolating Distance between Optimal Transport and Fischer-Rao, arXiv preprint arXiv:1506.06430, 2015.

Cordero-Erausquin, D., Nazaret, B., and Villani, C.: A mass-transportation approach to sharp Sobolev and Gagliardo–Nirenberg inequalities, Advances in Mathematics, 182, 307–332, 2004.

Cullen, M. and Gangbo, W.: A variational approach for the 2-dimensional semi-geostrophic shallow water equations, Archive for rational mechanics and analysis, 156, 241–273, 2001.

Delon, J. and Desolneux, A.: Stabilization of flicker-like effects in image sequences through local contrast correction, SIAM Journal on Imaging Sciences, 3, 703–734, 2010.

Farchi, A., Bocquet, M., Roustan, Y., Mathieu, A., and Quérel, A.: Using the Wasserstein distance to compare fields of pollutants: application to the radionuclide atmospheric dispersion of the Fukushima-Daiichi accident, Tellus B: Chemical and Physical Meteorology, 68, 31 682, doi:10.3402/tellusb.v68.31682, 2016.

Hamill, T. M. and Snyder, C.: A Hybrid Ensemble Kalman Filter-3D Variational Analysis Scheme, Monthly Weather Review, 128, 2905–2919, 2000.

Hoffman, R. N. and Grassotti, C.: A technique for assimilating SSM/I observations of marine atmospheric storms: tests with ECMWF analyses, Journal of Applied Meteorology, 35, 1177–1188, 1996.

Law, K., Stuart, A., and Zygalakis, K.: Data assimilation: a mathematical introduction, vol. 62, Springer, 2015.

Lewis, J. M., Lakshmivarahan, S., and Dhall, S.: Dynamic data assimilation: a least squares approach, vol. 13, Cambridge University Press, 2006.

Maury, B., Roudneff-Chupin, A., and Santambrogio, F.: A macroscopic crowd motion model of gradient flow type, Mathematical Models and Methods in Applied Sciences, 20, 1787–1821, 2010.

Mérigot, Q.: A multiscale approach to optimal transport, in: Computer Graphics Forum, vol. 30, pp. 1583–1592, Wiley Online Library, 2011.

Monge, G.: Mémoire sur la théorie des déblais et des remblais, De l'Imprimerie Royale, 1781.

Ning, L., Carli, F. P., Ebtehaj, A. M., Foufoula-Georgiou, E., and Georgiou, T. T.: Coping with model error in variational data assimilation using optimal mass transport, Water Resources Research, 50, 5817–5830, 2014.

Nocedal, J. and Wright, S. J.: Numerical Optimization, Springer series in Operations Research and Financial Engineering, Springer, 2006.

Otto, F.: The geometry of dissipative evolution equations: the porous medium equation, Communications in Partial Differential Equations, 26, 101–174, 2001.

Papadakis, N., Peyré, G., and Oudet, E.: Optimal transport with proximal splitting, SIAM Journal on Imaging Sciences, 7, 212–238, 2014.

Park, S. K. and Xu, L.: Data assimilation for atmospheric, oceanic and hydrologic applications, vol. 1, Springer Science & Business Media, 2009.

Park, S. K. and Xu, L.: Data assimilation for atmospheric, oceanic and hydrologic applications, vol. 2, Springer Science & Business Media, 2013.

Ratner, V., Zhu, L., Kolesov, I., Nedergaard, M., Benveniste, H., and Tannenbaum, A.: Optimal-mass-transfer-based estimation of glymphatic transport in living brain, in: SPIE Medical Imaging, vol. 9413, pp. 94 131J–94 131J–6, International Society for Optics and Photonics, 2015.

Ravela, S., Emanuel, K., and McLaughlin, D.: Data assimilation by field alignment, Physica D: Nonlinear Phenomena, 230, 127–145, 2007.

Rubner, Y., Tomasi, C., and Guibas, L. J.: A metric for distributions with applications to image databases, in: Computer Vision, 1998. Sixth International Conference on, pp. 59–66, IEEE, 1998.

Rubner, Y., Tomasi, C., and Guibas, L. J.: The earth mover's distance as a metric for image retrieval, International journal of computer vision, 40, 99–121, 2000.

Villani, C.: Topics in optimal transportation, no. 58 in Graduate studies in mathematics, American Mathematical Soc., 2003.