Optimal Transport for Data Assimilation

Nelson Feyeux, Maëlle Nodet, Arthur Vidard

Optimal Transport for Data Assimilation

Nelson Feyeux[1], Maelle Nodet[2], Arthur Vidard[1]
1: INRIA, 2: Univ. Grenoble Alpes

17th June 2016

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 and it shows 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 [LLD06, LSZ15]). 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 [PX09, PX13]).

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 [HS00, Bue05, BS14]. 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 [HG96, REM07]. 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.

The optimal transport theory has been founded by Monge in 1781 [Mon81]. 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

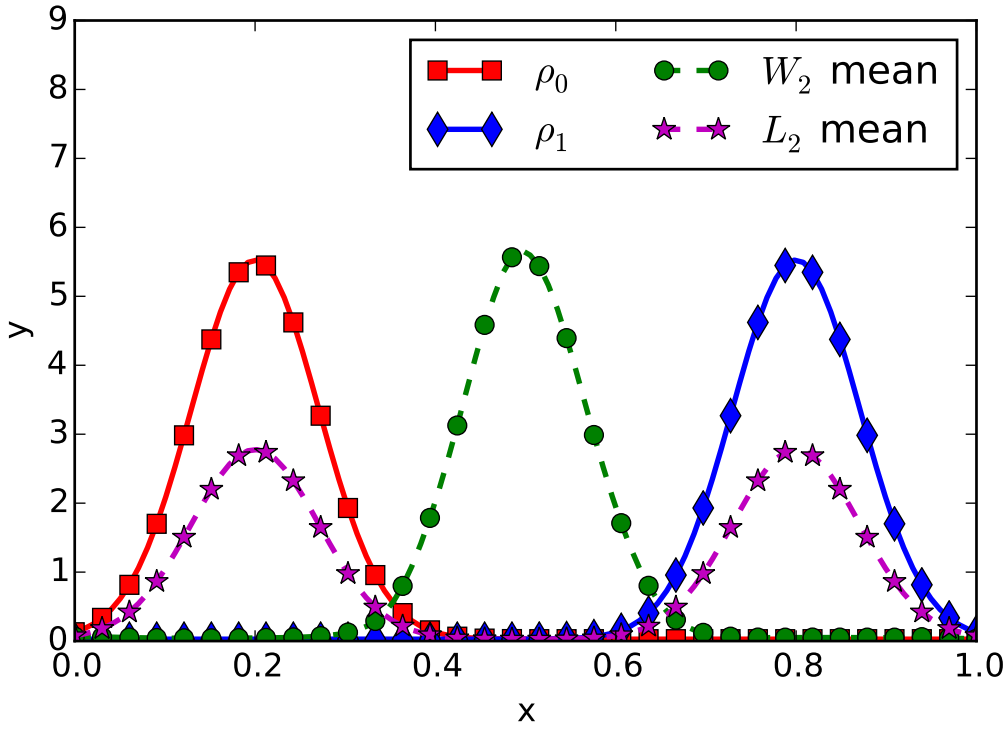


Figure 1: Wasserstein ($\mathcal{W}_2$) and Euclidean ($\mathcal{L}_2$) means of two densities ρ_0 and ρ_1.

as a transportation problem between two probability densities. A modern presentation can be found in [Vil03] 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 [CENV04] to the semi-geostrophic equations [CG01], through astrophysics [BFH+03], medicine [RZK+15], crowd motion [MRCS10] or urban planning [BS05]. 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 [RTG98, RTG00], interpolation [BVDPPH11], or movie reconstruction [DD10].

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 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. The related cost function is formulated and its minimization described in Section 3. Finally, in Section 4 numerical illustrations are presented, 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

$$\mathcal{J}(\mathbf{x}) = d_R(\mathcal{G}(\mathbf{x}), \mathbf{y}^{\text{obs}})^2 + d_B(\mathbf{x}, \mathbf{x}^b)^2, \tag{2.1}$$

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

$$\mathcal{J}(\mathbf{x}) = \|\mathcal{G}(\mathbf{x}) - \mathbf{y}^{\text{obs}}\|_2^2 + \|\mathbf{x} - \mathbf{x}^b\|_2^2, \tag{2.2}$$

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

$$\|\mathbf{x}\|_2^2 := \int |\mathbf{x}(x)|^2 \,\mathrm{d}x. \tag{2.3}$$

The Wasserstein distance $\mathcal{W}_2$ in place of d_R and d_B in equation (2.1) 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:

$$\mathcal{P}(\Omega) := \left\{ \rho \geq 0 \colon \int_\Omega \rho(x) \,\mathrm{d}x = 1 \right\}. \tag{2.4}$$

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,

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho \mathbf{v}) = 0. \tag{2.5}$$

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 $\rho(t,x)\mathbf{v}(t,x)\cdot\vec{n}(x)=0$ 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,

$$\mathcal{W}_2^2(\rho_0,\rho_1) = \min_{(\rho,\mathbf{v})\in C(\rho_0,\rho_1)} \iint_{[0,1]\times\Omega} \rho(t,x)|\mathbf{v}(t,x)|^2 \,\mathrm{d}t\mathrm{d}x \tag{2.6}$$

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,

$$C(\rho_0,\rho_1) := \left\{ (\rho,\mathbf{v}) \text{ s.t. } \begin{array}{l} \partial_t\rho + \operatorname{div}(\rho\mathbf{v}) = 0, \\ \rho(t=0)=\rho_0,\ \rho(t=1)=\rho_1, \\ \rho\mathbf{v}\cdot\vec{n} = 0 \text{ on } \partial\Omega \end{array} \right\}. \tag{2.7}$$

This definition of the Wasserstein distance is the Benamou-Brenier formulation [BB00]. 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 [Vil03] 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 [AGS08]

$$\partial_t\Phi + \frac{|\nabla\Phi|^2}{2} = 0. \tag{2.8}$$

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 [PPO14] or the semi-discrete [Mér11].

2.2.3 Wasserstein inner product

The scalar product between two functions is required for data assimilation and optimization. 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. [Ott01]),

$$T_\rho\mathcal{P} = \left\{ -\mathrm{div}(\rho\nabla\Phi) \in \mathcal{L}^2(\Omega), \rho\frac{\partial\Phi}{\partial\vec{n}} = 0 \text{ on } \partial\Omega \right\}. \tag{2.9}$$

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

$$\forall\eta,\eta' \in T_\rho\mathcal{P}(\Omega), \quad \langle\eta,\eta'\rangle_2 := \int_\Omega \eta(x)\eta'(x)\,\mathrm{d}x. \tag{2.10}$$

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

$$\langle\eta,\eta'\rangle_W := \int_\Omega \rho\,\nabla\Phi\cdot\nabla\Phi'\,\mathrm{d}x. \tag{2.11}$$

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

$$\|\eta\|_W^2 = \int_\Omega \rho|\nabla\Phi|^2 \tag{2.12}$$

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

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 [CSPV15]. For simplicity we do not consider here this possible generalization and all data have the same integral. The cost function (2.1) is rewritten using the Wasserstein distance defined in Section 2.2,

$$\mathcal{J}_{\mathcal{W}}(\mathbf{x}_0) = \frac{1}{2}\sum_{i=1}^{N^{\text{obs}}} \mathcal{W}_2^2(\mathcal{G}_i(\mathbf{x}_0), \mathbf{y}_i^{\text{obs}}) + \frac{\omega_b}{2}\mathcal{W}_2^2(\mathbf{x}_0, \mathbf{x}_0^b), \tag{3.1}$$

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 (3.1) then becomes

$$\mathcal{J}_{\mathcal{W}}(\rho_0) = \frac{1}{2}\sum_{i=1}^{N^{\text{obs}}} \mathcal{W}_2^2(\mathcal{G}_i(\rho_0), \rho_i^{\text{obs}}) \quad + \quad \frac{\omega_b}{2}\,\mathcal{W}_2^2(\rho_0, \rho_0^b). \tag{3.2}$$

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

$$\forall \eta \in T_{\rho_0}\mathcal{P}, \quad \lim_{\epsilon \to 0} \frac{\mathcal{J}_{\mathcal{W}}(\rho_0 + \epsilon\eta) - \mathcal{J}_{\mathcal{W}}(\rho_0)}{\epsilon} = \langle \eta, g \rangle \tag{3.3}$$

Chosing another scalar product will give another gradient. The choice of the scalar product / 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 $\mathrm{grad}_2\mathcal{J}_{\mathcal{W}}(\rho_0)$ and $\mathrm{grad}_W\mathcal{J}_{\mathcal{W}}(\rho_0)$ and are the only elements of the tangent space $T_{\rho_0}\mathcal{P}$ of $\rho_0 \in \mathcal{P}(\Omega)$ such that

$$\begin{aligned} \forall \eta \in T_{\rho_0}\mathcal{P}, \quad \lim_{\epsilon \to 0} \frac{\mathcal{J}_{\mathcal{W}}(\rho_0 + \epsilon\eta) - \mathcal{J}_{\mathcal{W}}(\rho_0)}{\epsilon} &= \langle \mathrm{grad}_2\mathcal{J}_{\mathcal{W}}(\rho_0), \eta \rangle_2 \\ &= \langle \mathrm{grad}_W\mathcal{J}_{\mathcal{W}}(\rho_0), \eta \rangle_W. \end{aligned} \tag{3.4}$$

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,*

$$\mathrm{grad}_2\mathcal{J}_{\mathcal{W}}(\rho_0) = \omega_b\Psi^b + \sum_{i=1}^{N^{obs}} \mathbf{G}_i^*(\rho_0).\Psi^i + c \tag{3.5}$$

with c such that the integral of $\mathrm{grad}_2\mathcal{J}_{\mathcal{W}}(\rho_0)$ 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 $\mathrm{grad}_2\mathcal{J}_{\mathcal{W}}(\rho_0)$ has the no-flux boundary condition (see comment about this assumption below)

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

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

$$\mathrm{grad}_W \mathcal{J}_{\mathcal{W}}(\rho_0) = -\mathrm{div}\Big(\rho_0 \nabla[\mathrm{grad}_2 \mathcal{J}_{\mathcal{W}}(\rho_0)]\Big). \tag{3.6}$$

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

$$\forall \eta, \mu \in T_{\rho_0}\mathcal{P}, \langle \mathbf{G}_i^*(\rho_0).\mu, \eta\rangle_2 = \langle \mu, \mathbf{G}_i(\rho_0).\eta\rangle_2 \tag{3.7}$$

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

$$\forall \eta \in T_{\rho_0}\mathcal{P}, \mathbf{G}_i(\rho_0).\eta := \lim_{\epsilon\to 0} \frac{\mathcal{G}_i(\rho_0 + \epsilon\eta) - \mathcal{G}_i(\rho_0)}{\epsilon}. \tag{3.8}$$

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

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

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 [CSPV15].

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

The minimizer of $\mathcal{J}_{\mathcal{W}}$ defined in (3.2) 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

$$\rho_0^{n+1} = \rho_0^n - \alpha^n d^n \tag{3.9}$$

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 [NW06] for more details.

Remark 3.4 (Note on descent with the Wasserstein gradient)**.** With the Wasserstein gradient (3.6), the descent of $\mathcal{J}_{\mathcal{W}}$ follows an iteration scheme of the form

$$\rho_0^{n+1} = \rho_0^n + \alpha^n \operatorname{div}(\rho_0^n \nabla \Phi^n). \tag{3.10}$$

A more transport-like iteration could be used instead,

$$\rho_0^{n+1} = (I - \alpha^n \nabla \Phi^n) \# \rho_0^n \tag{3.11}$$

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 (3.11) 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 (3.11) is equivalent to (3.10) 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

$$\begin{cases} \partial_t \rho + \operatorname{div}(\rho \nabla \Phi) = 0 \\ \partial_t \Phi + \dfrac{|\nabla \Phi|^2}{2} = 0. \end{cases} \tag{3.12}$$

with initial conditions $\rho(0,x) = \rho_0(x)$ and $\Phi(0,x) = \Phi_0(x)$, see [Vil03, (5.61)]. Therefore, (3.10) is just an explicit time-discretization of (3.12) and is equivalent to (3.11) 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.

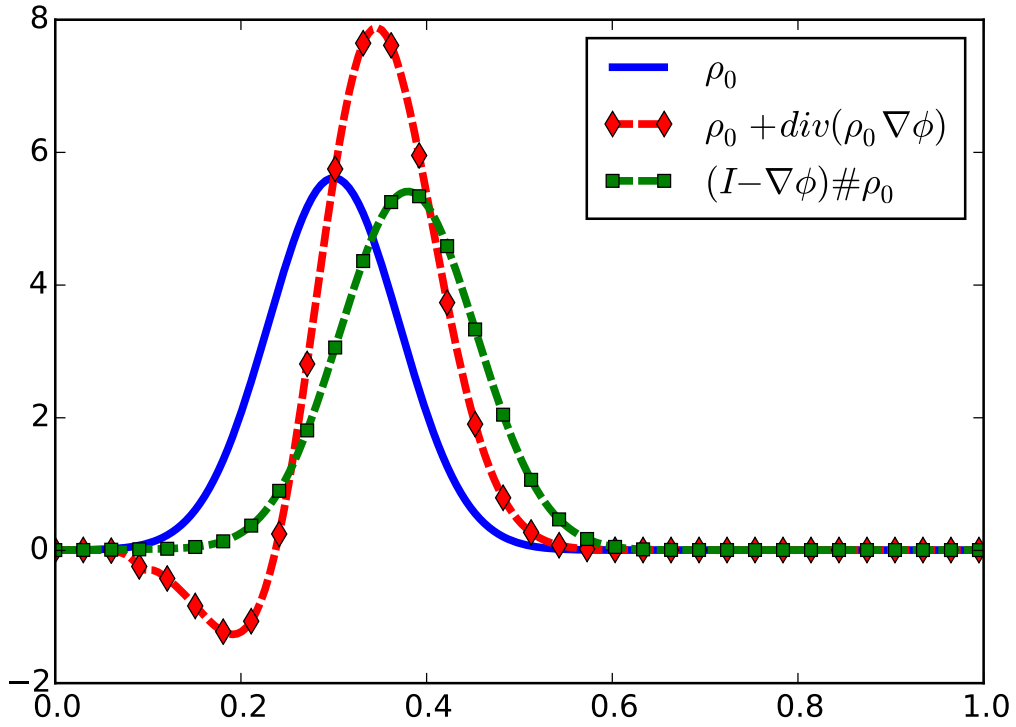


Figure 2: Comparison of iterations (3.10) and (3.11) with ρ_0 of limited support and Φ such that $\nabla\Phi$ is constant on the support of ρ_0.

In this section are presented the analyses resulting from the minimization of the Wasserstein cost function defined previously in (DG2), in particular when position errors occur. Results are compared with the results given by the $\mathcal{L}^2$ cost function defined in (2.2).

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 obervations 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

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

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

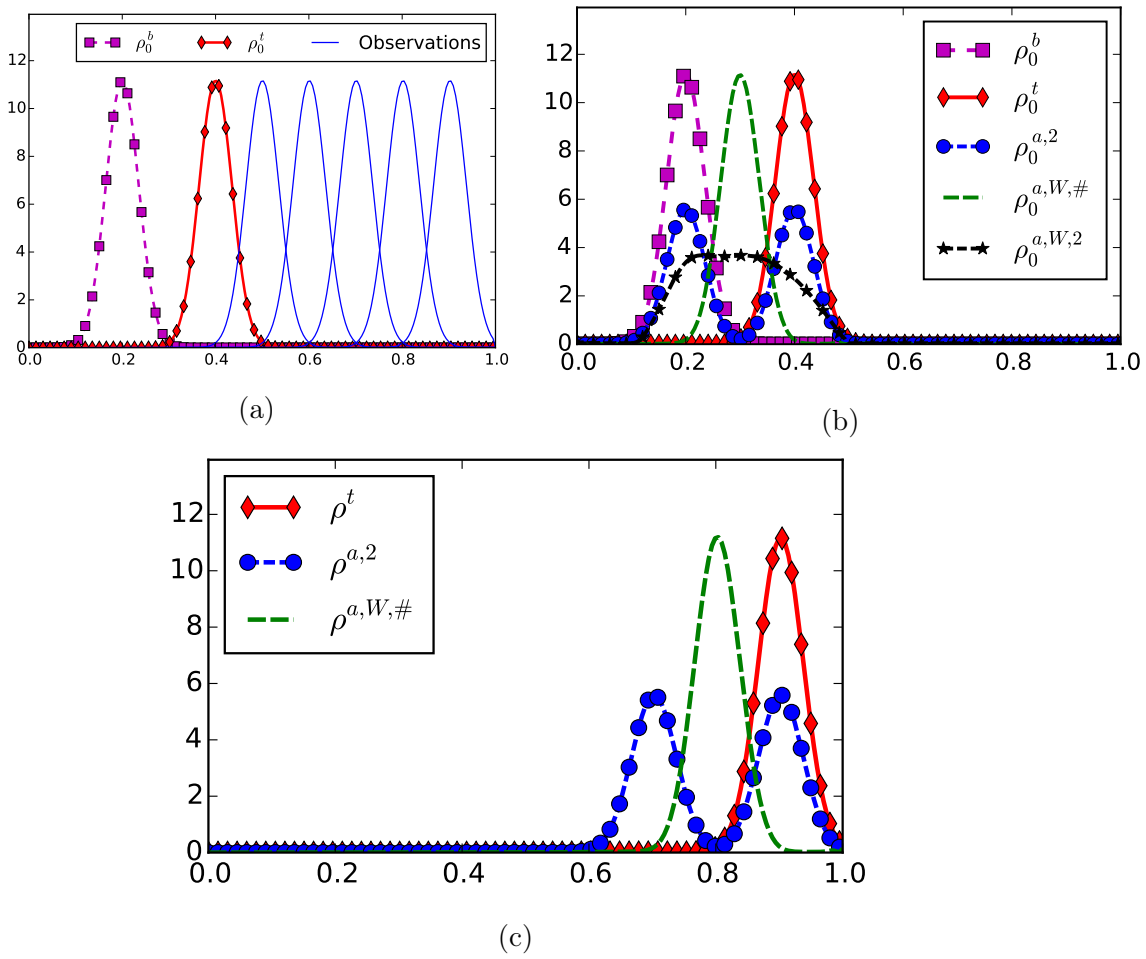
The value of α^n is chosen as optimal on each iteration and the algorithm stops when the decrement of $\mathcal{J}$ between two iterations is lower than 10^{-6}.

4.1 Linear example

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

$$\partial_t\rho + \nabla\rho = 0. \tag{4.2}$$

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) (b)

(c)

Figure 3: Plot (a) shows the twin experiments ingredients: true initial condition ρ_0^t, background term ρ_0^b, and observations at different times. Plot (b) shows 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#). In (c) 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.

a different place, as if it had position errors. On Fig. 3a 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. 3b. 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. 3c. At the end of the assimilation window, each of both of the analyses still have discrepancies with the observations.

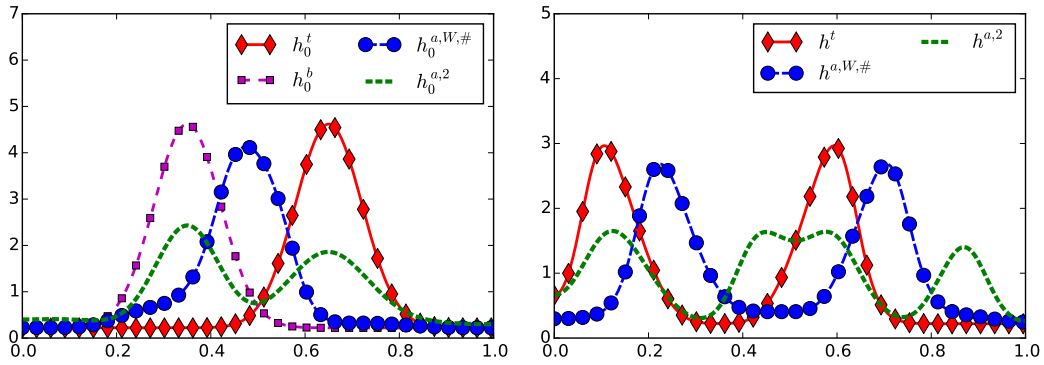
As a conclusion of this first test case, we managed to write and minimize a cost function which gives a relevant analysis, contrary to what we obtain with the classical Euclidean cost function, in case of position errors. We also noticed that the success of the minimization of $\mathcal{J}_{\mathcal{W}}$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



(a) (b)

Figure 4: On (a) 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. The Figure (b) shows the same plots (except the background one) but at the output of the model.

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 the observations 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. 4a 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).

The Fig. 4b 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.

Conclusion

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.

Acknowledgements

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

A Proof of Theorem 3.1

To prove Theorem 3.1, one first needs to differentiate the Wasserstein distance. The following Lemma from [Vil03, 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* $\epsilon \in \mathbb{R}$*,*

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

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 (3.1), from the defintion of the tangent model (3.8) and in application of the Lemma A.1,

$$\begin{aligned}
\lim_{\epsilon \to 0} \frac{\mathcal{J}_{\mathcal{W}}(\rho_0 + \epsilon\eta) - \mathcal{J}_{\mathcal{W}}(\rho_0)}{\epsilon} &= \sum_{i=1}^{N^{obs}} \langle \mathbf{G}_i[\rho_0]\eta, \phi^i \rangle_2 + \omega_b \langle \eta, \phi^b \rangle_2 \\
&= \left\langle \eta, \sum_{i=1}^{N^{obs}} \mathbf{G}_i^*[\rho_0]\phi^i + \omega_b \phi^b \right\rangle_2 \\
&= \left\langle \eta, \sum_{i=1}^{N^{obs}} \mathbf{G}_i^*[\rho_0]\phi^i + \omega_b \phi^b + c \right\rangle_2
\end{aligned} \tag{A.2}$$

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

$$\mathrm{grad}_2 \mathcal{J}_{\mathcal{W}}(\rho_0) = \sum_{i=1}^{N^{obs}} \mathbf{G}_i^*[\rho_0]\phi^i + \omega_b \phi^b + c \tag{A.3}$$

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 (A.2) and (A.3) give

$$\begin{aligned}
\langle \eta, g \rangle_2 &= \langle -\mathrm{div}(\rho_0 \nabla \Phi), g \rangle_2 \\
&= -\int_\Omega \mathrm{div}(\rho_0 \nabla \Phi) g \\
&= \int_\Omega \rho_0 \nabla \Phi \nabla g.
\end{aligned} \tag{A.4}$$

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

$$\int_\Omega \rho_0 \nabla \Phi \nabla g = \langle \eta, -\mathrm{div}(\rho_0 \nabla g) \rangle_W, \tag{A.5}$$

hence

$$\forall \eta \in T_{\rho_0}\mathcal{P}, \quad \lim_{\epsilon \to 0} \frac{\mathcal{J}_{\mathcal{W}}(\rho_0 + \epsilon\eta) - \mathcal{J}_{\mathcal{W}}(\rho_0)}{\epsilon} = \langle \eta, -\mathrm{div}(\rho_0 \nabla g) \rangle_W. \tag{A.6}$$

□

References

[AGS08] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. Gradient flows: in metric spaces and in the space of probability measures. Springer Science & Business Media, 2008.

[BB00] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. Numerische Mathematik, 84(3):375–393, 2000.

[BFH+03] Yann Brenier, Uriel Frisch, Michel Hénon, Grégoire Loeper, Sabino Matarrese, Roya Mohayaee, and Andrei Sobolevskiĭ. Reconstruction of the early universe as a convex optimization problem. Monthly Notices of the Royal Astronomical Society, 346(2):501–524, 2003.

[BS05] Giuseppe Buttazzo and Filippo Santambrogio. A model for the optimal planning of an urban area. SIAM Journal on Mathematical Analysis, 37(2):514–530, 2005.

[BS14] M. Bocquet and P. Sakov. An iterative ensemble Kalman smoother. Quaterly Journal of the Royal Meteorological Society, 140:1521–1535, 2014.

[Bue05] M. Buehner. 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, 2005.

[BVDPPH11] Nicolas Bonneel, Michiel Van De Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement interpolation using Lagrangian mass transport. In ACM Transactions on Graphics (TOG), volume 30, page 158. ACM, 2011.

[CENV04] Dario Cordero-Erausquin, Bruno Nazaret, and Cédric Villani. A mass-transportation approach to sharp Sobolev and Gagliardo–Nirenberg inequalities. Advances in Mathematics, 182(2):307–332, 2004.

[CG01] Mike Cullen and Wilfrid Gangbo. A variational approach for the 2-dimensional semi-geostrophic shallow water equations. Archive for rational mechanics and analysis, 156(3):241–273, 2001.

[CSPV15] Lenaic Chizat, Bernhard Schmitzer, Gabriel Peyré, and François-Xavier Vialard. An interpolating distance between optimal transport and fischer-rao. arXiv preprint arXiv:1506.06430, 2015.

[DD10] Julie Delon and Agnes Desolneux. Stabilization of flicker-like effects in image sequences through local contrast correction. SIAM Journal on Imaging Sciences, 3(4):703–734, 2010.

[HG96] Ross N Hoffman and Christopher Grassotti. A technique for assimilating SSM/I observations of marine atmospheric storms: tests with ECMWF analyses. Journal of Applied Meteorology, 35(8):1177–1188, 1996.

[HS00] T. M. Hamill and C. Snyder. A hybrid ensemble Kalman filter-3D variational analysis scheme. Monthly Weather Review, 128:2905–2919, 2000.

[LLD06] John M Lewis, Sivaramakrishnan Lakshmivarahan, and Sudarshan Dhall. Dynamic data assimilation: a least squares approach, volume 13. Cambridge University Press, 2006.

[LSZ15] Kody Law, Andrew Stuart, and Konstantinos Zygalakis. Data assimilation: a mathematical introduction, volume 62. Springer, 2015.

[Mér11] Quentin Mérigot. A multiscale approach to optimal transport. In Computer Graphics Forum, volume 30, pages 1583–1592. Wiley Online Library, 2011.

[Mon81] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. De l'Imprimerie Royale, 1781.

[MRCS10] Bertrand Maury, Aude Roudneff-Chupin, and Filippo Santambrogio. A macroscopic crowd motion model of gradient flow

type. Mathematical Models and Methods in Applied Sciences, 20(10):1787–1821, 2010.

[NW06] Jorge Nocedal and Stephen J. Wright. Numerical Optimization. Springer series in Operations Research and Financial Engineering. Springer, 2006.

[Ott01] Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. Communications in Partial Differential Equations, 26:101–174, 2001.

[PPO14] Nicolas Papadakis, Gabriel Peyré, and Edouard Oudet. Optimal transport with proximal splitting. SIAM Journal on Imaging Sciences, 7(1):212–238, 2014.

[PX09] Seon Ki Park and Liang Xu. Data assimilation for atmospheric, oceanic and hydrologic applications, volume 1. Springer Science & Business Media, 2009.

[PX13] Seon Ki Park and Liang Xu. Data assimilation for atmospheric, oceanic and hydrologic applications, volume 2. Springer Science & Business Media, 2013.

[REM07] Sai Ravela, Kerry Emanuel, and Dennis McLaughlin. Data assimilation by field alignment. Physica D: Nonlinear Phenomena, 230(1):127–145, 2007.

[RTG98] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. A metric for distributions with applications to image databases. In Computer Vision, 1998. Sixth International Conference on, pages 59–66. IEEE, 1998.

[RTG00] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. International journal of computer vision, 40(2):99–121, 2000.

[RZK+15] Vadim Ratner, Liangjia Zhu, Ivan Kolesov, Maiken Nedergaard, Helene Benveniste, and Allen Tannenbaum. Optimal-mass-transfer-based estimation of glymphatic transport in living brain. In SPIE Medical Imaging, volume 9413,

pages 94131J–94131J–6. International Society for Optics and Photonics, 2015.

[Vil03] C. Villani. Topics in Optimal Transportation. Graduate studies in mathematics. American Mathematical Society, 2003.