Coping with model error in variational data assimilation using optimal mass transport

Lipeng Ning,1 Francesca P. Carli,1 Ardeshir Mohammad Ebtehaj,2 Efi Foufoula-Georgiou,2 Tryphon T. Georgiou,1

1Department of Electrical and Computer Engineering, University of Minnesota, 200 Union Street SE, Minneapolis, MN, USA
2Department of Civil Engineering, St. Anthony Falls Laboratory, and National Center for Earth-Surface Dynamics, University of Minnesota, 2 Third Avenue SE, Minneapolis, MN, USA
Abstract. Classical variational data assimilation methods address the problem of optimally combining model predictions with observations in the presence of zero-mean Gaussian random errors. However, in many natural systems, uncertainty in model structure and/or model parameters often results in systematic errors or biases. Prior knowledge about such systematic model error for parametric removal is not always feasible in practice, limiting the efficient use of observations for improved prediction. The main contribution of this work is to advocate the relevance of transportation metrics for quantifying non-random model error in variational data assimilation for non-negative natural states and fluxes. Transportation metrics (also known as Wasserstein metrics) originate in the theory of Optimal Mass Transport (OMT) and provide a non-parametric way to compare distributions which is natural in the sense that it penalizes mismatch in the values and relative position of “masses” in the two distributions. We demonstrate the promise of the proposed methodology using 1D and 2D advection-diffusion dynamics with systematic error in the velocity and diffusivity parameters. Moreover, we combine this methodology with additional regularization functionals, such as the $\ell_1$-norm of the state in a properly chosen domain, to incorporate both model error and potential prior information in the presence of sparsity or sharp fronts in the underlying state of interest.
1. Introduction

Data assimilation aims at estimating the state of a physical system based on time-distributed observations, a dynamical model describing the space-time evolution of the underlying state, and a possible piece of prior information on the initial condition. In hydrologic and atmospheric systems, variational data assimilation is now a necessary component of any operational forecast system [see, Kalnay et al., 2007]. It is also an essential element of re-analysis methods which aim at producing physically consistent historical records for applications related to identifying climatic trends, testing and calibrating climate models, and detecting regional changes in moisture, precipitation, and temperature [see, Bengtsson et al., 2004]. From the statistical point of view, classic formulations of variational data assimilation methods rely on the assumption of *random errors* in observations and models. However, in reality most variational data assimilation systems are affected by *systematic errors*. By systematic errors, we refer to biases and all structured deviations of the model predictions from the true state. In these cases, even when effort is made to remove biases, the presence of residual biases prevents optimal usage of the available data [Dee, 2005].

In the classic 3D and 4D variational data assimilation methods, typically a quadratic functional is defined which encodes the weighted Euclidean distance of the true initial state to the background and observation states, while the best estimate of the initial state is its stationary point. Using variational calculus for data assimilation problems traces back to the pioneering work of Sasaki [1970], Lorenc [1986], Courtier and Talagrand [1990], among others. Nowadays, variational data assimilation methodologies are at the core of
atmospheric prediction [e.g., Lorenc et al., 2000; Kleist et al., 2009; Johnson et al., 2005; Kalnay et al., 2007; Houtekamer et al., 2005], soil moisture, land surface flux estimation [e.g., Crow and Wood, 2003; Caparrini et al., 2004; Houser et al., 1998; Reichle et al., 2008; Sini et al., 2008; Kumar et al., 2008; Bateni and Entekhabi, 2012], and hydrologic forecasting [McLaughlin, 2002; Vrugt et al., 2005; Liu and Gupta, 2007; Durand and Margulis, 2007; Hendricks Franssen and Kinzelbach, 2008; Moradkhani et al., 2012] – see also overviews by Reichle et al. [2002] and Liu et al. [2012].

There is a large body of research in variational data assimilation devoted to accounting for the characterization of systematic and random model errors. For instance, the problem of accounting for bias has been tackled by considering serially correlated random error [e.g., Lorenc, 1986; Derber, 1989; Zupanski, 1997a] while Ghahramani and Roweis [1999] introduced a semi-parametric framework. Griffith and Nichols [2000] presented a general parametric framework to treat the model error in the context of a non-zero mean Gaussian process, which has been further explored in Martin et al. [2002] for bias treatment in oceanic data assimilation problems. In this approach, simple assumptions about the evolution of the error are made, enabling the systematic error to be taken into account in the standard formulation of variational data assimilation [see, Nichols, 2003]. Also, Dee [2005] argued that one of the easiest ways to detect and treat model bias is to evaluate whether the behavior of the analysis has a tendency to make systematic corrections to the model background. Consequently, Dee [2005] suggested introducing an augmented state vector which includes bias terms as a set of constants that can be estimated by finding the stationary point of the augmented variational cost function in a semi-supervised fashion.
Model error in a variational data assimilation framework will typically lead to an incorrect distribution of the predicted states. Discrepancies between these predicted states and the (unknown) true states might include a translational bias, an incorrect spread of the mass, or higher-order distortion of the entire distribution, all considered as special cases of what we call systematic model errors. The central focus of this paper is to address the question as to whether there is a distance metric that can naturally characterize the mismatch between model predicted and true states and non-parametrically account for systematic errors. We propose that the optimal mass transportation (OMT) metric (or Wasserstein metric) serves as a natural metric to quantify such systematic errors since it captures and rewards the “similarity” between two distributions with respect to the values and relative position of the “mass” of the distributions (see also Rubner et al. [2000] for a complementary viewpoint on why the OMT metric is natural in this context). By incorporating this metric within the classical variational data assimilation framework, we present a methodology for obtaining improved estimates of the states in the presence of systematic errors.

This paper is structured as follows. In Section 2, classical formulations of variational data assimilation are reviewed. Section 3 explains the relevant theory of OMT while the essence of this metric for model error characterization is presented via a simple example. Section 4 presents the main contribution of this paper and incorporates the OMT-metric in the formulation of the variational data assimilation problem. Section 5 focuses on advection-diffusion dynamics for which we can explicitly demonstrate that the OMT-metric is a continuous function of the error in the space of model parameters and derive an upper bound on the distance between modeled and true states. Detailed examples
that show the effectiveness and potential of the proposed methodology are also reported in this section. Finally, conclusions are drawn in Section 6.

2. Variational data assimilation

Let us denote the dynamical evolution of the state space and the observation model as follows:

\[ x_{k+1} = F_k(x_k) + w_k \]  
\[ y_k = H_k(x_k) + v_k, \]

where \( F(\cdot), H(\cdot) \) denote the model and observation operators, respectively, and the (column) vectors \( w, v \) represent process and observation noise. A prior estimate of the true initial state \( x_0 \) in data assimilation is typically referred to as the “background state” \( x_b \), which is often obtained from a previous-time forecast or climatological information. Then, starting with \( K + 1 \) sequential measurements \( y_0, y_1, \ldots, y_K \), the problem is to estimate the states \( x_0, x_1, \ldots, x_K \). In the statistical filtering theory, this problem is referred to as a fixed-interval smoothing problem. If the error components \( w_k, v_k \ (k \in \{0, \ldots, K\}) \) and \( x_0 - x_0 \) are independent zero-mean Gaussian random variables while \( F_k(\cdot) \) and \( H_k(\cdot) \) are both linear time-varying operators, the solution is provided by the well-known Kalman filter [Kalman, 1960]. In the data assimilation literature, this problem is typically called 4-dimensional variational (4D-Var) data assimilation [see, e.g., Trémolet, 2006; Sasaki, 1970] and its equivalence to Kalman smoothing has been pointed out by Fisher et al. [2005]. In the classic formulation of the 4D-Var problem for deterministic dynamics with no model error \( w \), estimation of the underlying states amounts to minimizing the following
where $P$ and $R_k$ represent the covariances of the background error $e := x_0 - x_b$ and $v_k$, respectively. Here, the notation $\|e\|^2_{P_{-1}}$ represents the weighted quadratic norm, that is $e^T P^{-1} e$, where $T$ is the transpose operator. In order to account for model error in variational data assimilation, several techniques have been proposed [Derber and Rosati, 1989; Zupanski, 1997b; Vidard et al., 2004; Trémolet, 2006, among many others]. Typically, a stochastic model error $w$ is incorporated which leads to the following problem formulation, which is often called weak constraint 4D-Var:

$$\min_{x_0, \ldots, x_K} \left\{ \|x_0 - x_b\|^2_{P_{-1}} + \sum_{k=0}^{K} \|y_k - H_k(x_k)\|^2_{R_{k-1}} + \sum_{k=1}^{K} \|x_k - F_{k-1}(x_{k-1})\|^2_{Q_{k-1}} + \gamma \|\Phi(x_0)\|_1 \right\},$$

(3)

where similarly $Q_k$ represents the covariance of $w_k$.

More recently, the use of $\ell_1$-regularization in data assimilation was suggested by Freitag et al. [2012], Ebtehaj and Foufoula-Georgiou [2013] and Ebtehaj et al. [2014], to account for distinct geometrical features and the singular structure of the underlying state such as ridges and isolated jumps (e.g., sharp weather fronts or extreme rain cells) as well as sparsity of the underlying state in suitable transform domains. For instance, if it is known that $x_0$ is “smooth,” except for a few distinct jumps or that the derivative of $x_0$ has a Laplace-like distribution, then the $\ell_1$-norm $\|\Phi(x_0)\|_1$ with $\Phi$ a linear derivative-like operator (e.g., wavelet transform) can be used as a regularization term as follows:

$$\min_{x_0, \ldots, x_K} \left\{ \|x_0 - x_b\|^2_{P_{-1}} + \sum_{k=0}^{K} \|y_k - H_k(x_k)\|^2_{R_{k-1}} + \sum_{k=1}^{K} \|x_k - F_{k-1}(x_{k-1})\|^2_{Q_{k-1}} + \gamma \|\Phi(x_0)\|_1 \right\},$$

(4)
where $\gamma$ is a non-negative constant maintaining a balance between goodness of fit to the available information (model output and observations) and the underlying regularity of the initial state.

3. Monge-Kantorovich Optimal Mass Transport (OMT)

The main drawback of the above formalisms of the variational data assimilation problem is that we are implicitly attributing the model error entirely to a stochastic (Gaussian) noise $w$. This attribution is not typically consistent with the commonly observed structural error in physically-based environmental models, which may deteriorate the quality of analysis and forecast skills of data assimilation systems. The main contribution of this work is to advocate the relevance of transportation metrics for quantifying model error in the variational data assimilation framework. These types of metrics are based on the theory of Monge-Kantorovich optimal mass transport (OMT) [see, Villani, 2003]. In particular, using transportation metrics in data assimilation problems allows us to naturally characterize the distance between the state of the model forecast and the unknown true state in a non-parametric fashion, without requiring any prior assumption about the model error which might be physically unrealistic or practically prohibitive.

In the following, we confine our discussion to the quadratic form of the OMT-metric (also known as Monge-Kantorovich problem of exponent 2) in discrete space and specialize its definition for variational data assimilation problems. To this end, let us consider two $n$-element density vectors $\mathbf{x}$ and $\tilde{\mathbf{x}}$, having a support on a discrete set of points $\{x_i\}_{i=1}^N$ and $\{\tilde{x}_i\}_{i=1}^N$, respectively. Usually the density vectors are obtained from discretization of continuous densities and the support sets correspond to space-time locations that are dictated by resolution and sampling rates. For now, assume that $\mathbf{x}$ and $\tilde{\mathbf{x}}$ have equal
mass, i.e. $\sum_{i=1}^{N} x(x_i) = \sum_{i=1}^{N} \tilde{x}(\tilde{x}_i)$, and the cost of transporting one unit of mass from location $x_i$ to $\tilde{x}_j$ is denoted by $c_{i,j}$. The Monge-Kantorovich mass transport problem considers the minimum cost of transferring all the mass from the distribution $x$ to the distribution $\tilde{x}$. The original formulation goes back to the work by Monge [1781], while the modern formulation is due to Kantorovich [1942]. Note that, this formulation casts the problem as a linear programming which can be solved efficiently in large dimensions.

More specifically, if $m(x_i, \tilde{x}_j)$ denotes the mass that is to be transported from location $x_i$ to location $\tilde{x}_j$, the OMT-metric $T(x, \tilde{x})$ can be computed as follows:

$$ T(x, \tilde{x}) := \min_{m} \sum_{i,j} c_{i,j} m(x_i, \tilde{x}_j) $$
subject to
$$ m(x_i, \tilde{x}_j) \geq 0, $$
$$ \sum_j m(x_i, \tilde{x}_j) = x(x_i), $$
$$ \sum_i m(x_i, \tilde{x}_j) = \tilde{x}(\tilde{x}_j). $$ (5)

The above formulation can be expressed more compactly if we let $1$ denote an $N$-dimensional vector with all the entries equal to one, $C$ denote an $N \times N$ matrix with the $(i, j)$th entry $c_{i,j}$ and, $M$ denote an $N \times N$ matrix with $M(i, j) = m(x_i, \tilde{x}_j)$. Then,

$$ T(x, \tilde{x}) := \min_{M} \text{tr}(CM) $$
subject to
$$ M1 = x, \ M^T1 = \tilde{x}, \ M(i, j) \geq 0 \ \forall i, j. $$ (6)

Here, the matrix $M$ is a joint density matrix with positive elements, often referred to as the transportation plan [Villani, 2003], where $x$ and $\tilde{x}$ are the marginal mass functions.

Note that for the rest of the paper we specialize our consideration to $C(i, j) = \|x_i - \tilde{x}_j\|^2$ for which the transportation cost is characterized by the Euclidean distance between the location of masses in $x_i$ and $\tilde{x}_i$. This selection of the quadratic transportation cost,
known as the OMT of exponent 2, allows us to remain in the domain of smooth and convex optimization.

The above definition of the OMT-metric assumes that the total mass of the state of interest is fully conserved. This assumption might be restrictive in data assimilation of non-conservative states, as for example in the presence of source and sink elements in the underlying dynamics. To this end, a more relaxed OMT formulation is required.

Different methods have been proposed for generalizing the OMT-metric to account for non-equal masses [see, e.g., Benamou, 2003; Georgiou et al., 2009]. For example, the method proposed in Benamou [2003] was to use a mixture of the OMT-metric and $\ell_2$-norm as follows:

$$\mathcal{T}_\sigma(x, \tilde{x}) := \min_{\hat{x}} \left\{ \mathcal{T}(\hat{x}, \tilde{x}) + \sigma \|x - \hat{x}\|_2^2 \right\},$$

where the non-negative parameter $\sigma$ is used to denote the relative significance of the $\ell_2$-cost. Notice that the above relaxed OMT metric is obtained through a nested minimization problem based on an intermediate state $\hat{x}$ with minimal $\ell_2$ distance to the true state $x$ (second term within the bracket), while at the same time its mass is considered to be equal to $\tilde{x}$, through the use of the classic $\mathcal{T}(\hat{x}, \tilde{x})$ metric (first term within the bracket). In other words, this new relaxed OMT metric penalizes the distance of $\hat{x}$ from the model predicted state $\tilde{x}$ in the OMT sense, while it keeps the solution close enough to the true state $x$ without strictly enforcing mass equality due to the presence of model error. Obviously, the parameter $\sigma$ accommodates this relaxation.

Before we proceed with formal technical details and more rigorous exposition of the use of OMT in variational data assimilation, we deem necessary to present a simple and intuitive example that illustrates the difference between $\mathcal{T}(x, \tilde{x})$ and $\|x - \tilde{x}\|_2$ and elaborate
on the essence of utilizing the transportation metric for exhaustive quantification of the model error. We also provide insight via an example on how naturally OMT reconciles two (oppositely) biased estimates. More specifically, when biases are in opposing directions, averaging in a quadratic sense will necessarily distort the original shapes while OMT preserves them.

Consider that the state $x_k(x)$ at time $k$ represents a density function with a compact support in a bounded set $\mathcal{D}$ where $x_k(x) > 0$ for $\forall x \in \mathcal{D}$. To this end, consider a simple non-diffusive advective state space as follows:

$$x_{k+1}(x) = x_k(x - d),$$

in which, for example we consider a drift value of $d = 3$ representing a velocity vector field. Thus, comparing to (1), $w$ is absent while $F$ is a constant shift matrix. For example, let the initial state be

$$x_0 = \begin{cases} 
10 & \text{for } -0.5 \leq x \leq 0.5 \\
0 & \text{otherwise},
\end{cases}$$

which is shown in the black solid line in Figure 1, while the state $x_1(x)$ is shown in the red solid line. To elaborate on the advantages of the OMT-metric for bias correction purposes, let us now consider four biased model outputs (estimates) of the above state space (ground truth) obtained by the erroneous advection parameters $d \in \{3.5, 4, 5.5, 7\}$, respectively. The corresponding estimates $\{\tilde{x}_1^{(i)}\}_{i=1}^4$ and a comparison of their distance from the true state based on the metrics $\|\cdot\|_2^2$ and $\mathcal{T}(\cdot)$ are shown in Figure 1. It can be seen that when the support sets of $x_1$ and $\tilde{x}_1$ are not overlapping, $\|x_1 - \tilde{x}_1\|_2^2$ is constant and independent of the bias magnitude whereas the transportation distance increases monotonically. In other words, the OMT-metric penalizes larger biases monotonically in this case, while the
$\ell_2$-cost is insensitive to the bias magnitudes after a certain threshold, depending on the structure of the support sets.

To further exemplify differences between the quadratic and OMT metrics we consider an (academic) example of reconciling two oppositely biased densities in Figure 2. The left panel displays the true density $x$ and two biased and noisy estimates on either side. The right panel displays the reconciled estimate of $x$ obtained via averaging the two biased states using the $\ell_2$ norm, $\hat{x}^{\ell_2}$, and using an OMT metric, $\hat{x}^{\text{omt}}$, and compares these with the true density $x$. In essence, the reconciled estimate represents a “mean” (in the $\ell_2$-sense and in the OMT-sense, respectively). Naturally, the $\ell_2$-estimate is bi-modal (whereas, the original density was not). As is evident, the OMT-estimate has a tendency to preserve and reconcile the shapes of the two biased densities.

4. OMT in variational data assimilation

We denote the transportation distance between the state $x_k$ and the predicted state $F_{k-1}(x_{k-1})$ as $T(x_k, F_{k-1}(x_{k-1}))$ and, likewise, the transportation distance between the true and background states as $T(x_0, x_b)$. Our proposed formulation of the variational data assimilation problem is as follows:

$$\min_{x_0, \ldots, x_K} \left\{ T(x_0, x_b) + \sum_{k=0}^{K} \| y_k - H_k(x_k) \|_{\ell_2}^2 + \sum_{k=1}^{K} T(x_k, F_{k-1}(x_{k-1})) \right\} \quad (8)$$

where the $\ell_2$-norm is replaced with the OMT metric for taking into account both systematic and random model errors. Here, the term corresponding to the measurement error is quantified by a quadratic norm under the assumption that observational errors are well represented by additive (Gaussian) random noise.
For the purposes of this paper, we specialize to the case where the model $F_k$ and the observation operator $H_k$ are linear operators. A refinement of (8), where we have replaced the $T$’s by $T_\sigma$’s to allow for the possibility of non-conservative states, is as follows:

$$\min_{x_0, \ldots, x_K} \left\{ T_\sigma(x_0, x_b) + \sum_{k=0}^{K} \| y_k - H_k x_k \|^2_{R_k^{-1}} + \sum_{k=1}^{K} T_\sigma(x_k, F_{k-1} x_{k-1}) + \gamma \| \Phi x_0 \|_1 \right\}.$$  

This includes also a regularization term $\| \Phi x_0 \|_1$ that could be used to promote particular features in the state (e.g., see Freitag et al. [2012] and Ebtehaj et al. [2014]). In view of (7), the optimization problem (8) can be written as

$$\min_{x_0, \ldots, x_K} \left\{ \sigma \left( \| x_0 - \bar{x}_b \|^2 + \sum_{k=1}^{K} \| x_k - \bar{x}_k \|^2 \right) + \sum_{k=0}^{K} \| y_k - H_k x_k \|^2_{R_k^{-1}} + \gamma \| \Phi x_0 \|_1 + T(\bar{x}_b, x_b) + \sum_{k=1}^{K} T(\bar{x}_k, F_{k-1} x_{k-1}) \right\}$$

where, in general, the terms $T(\bar{x}_b, x_b) + \sum_{k=1}^{K} T(\bar{x}_k, F_{k-1} x_{k-1})$ account for systematic model error (bias) of the states as explained before. Notice that $T_\sigma$ is itself the solution of an optimization problem in equation (7) and thus $\bar{x}_b, \bar{x}_1, \ldots, \bar{x}_K$ have been added as optimization variables.

Recalling that the OMT-metric is obtained by minimizing over the joint density matrix, the last term in (9) can be expanded as follows:

$$T(\bar{x}_k, F_{k-1} x_{k-1}) = \min_{M_k} \text{tr}(CM_k)$$
subject to $M_k(i, j) \geq 0$, $M_k 1 = \bar{x}_k$, $M_k^T 1 = F_{k-1} x_{k-1}$.

Thus the problem (9) can be comprehensively represented as follows:

$$\min_{x_0, \ldots, x_K} \left\{ \sigma \left( \| x_0 - M_b 1 \|^2 + \sum_{k=1}^{K} \| x_k - M_k 1 \|^2 \right) + \sum_{k=0}^{K} \| y_k - H_k x_k \|^2_{R_k^{-1}} + \text{tr} \left( C(M_b + \sum_{k=1}^{K} M_k) \right) + \gamma \| \Phi x_0 \|_1 \right\}$$
subject to $M_b(i, j), M_k(i, j) \geq 0$, $M_b 1 = x_0$, $M_k^T 1 = F_{k-1} x_{k-1}$, $k = 1, \ldots, K$. ```
where we replaced \( \hat{x}_b \) by \( M_b \mathbf{1} \) and \( \hat{x}_k \) by \( M_k \mathbf{1} \). We note that the size of \( M_k \)’s is \( N \times N \) since \( x_k \)’s are \( N \) dimensional states. Clearly, the computational complexity of this optimization problem is higher than that of the classic 4D-Var method (see comparison in the examples presented below). However, more efficient algorithms for solving large-scale OMT problems are being developed [Haker et al., 2004; Haber et al., 2010] and their adaptation to variational data assimilation problems will be the subject of future research.

5. Applications using the advection–diffusion equation

In this section we elaborate on the advantages of the proposed variational data assimilation focusing on the advection–diffusion model, which forms the basis of many geophysical models in atmospheric, oceanic, hydrologic, and land-surface data assimilation applications. In this model, we specifically explore the suitability of the OMT-metric to quantify structural model error due to uncertainty in model parameters including the advection velocity and diffusivity coefficient. We show that the OMT-metric is a continuous function of the model error and admits well defined upper bounds expressed as functions of the model parameters.

5.1. Quantifying advection–diffusion model error via OMT

Consider the one–dimensional-advection diffusion equation

\[
\partial_t x + u \partial_x x = D \partial_{xx} x
\]

where \( u \) and \( D \) denote the constant advection and diffusion rate, respectively. It is well known that the advection–diffusion equation transfers a point mass distribution to a Gaussian density with mean \( ut \) and variance \( 2Dt \). In particular, if the initial state is the
Dirac delta $\delta(x)$, then the solution of (10) is

$$x(t, x) = \frac{1}{\sqrt{4\pi Dt}} e^{-\frac{(x-ut)^2}{4Dt}}$$

which is a Gaussian distribution with mean $ut$ and variance $2Dt$ [see e.g., Rubin and Atkinson, 2001]. We denote the fundamental kernel of a general advection-diffusion as follows:

$$P_{ut, Dt}(x) := \frac{1}{\sqrt{4\pi Dt}} e^{-\frac{(x-ut)^2}{4Dt}}.$$ 

We also use the notation $P_{ut, Dt}$ or $P_{ut, Dt}(\cdot)$ to denote this function. Given an arbitrary initial state $x_0$, the advection–diffusion equation (10) is linear in $x$, with the following solution:

$$x(t, x) = \int P_{ut, Dt}(x-y)x_0(y)dy.$$ 

(11)

Now, let us consider the advection–diffusion equations

$$\partial_t x + u_i \partial_x x = D_i \partial_{xx} x, \text{ for } i = p, q,$$

(12)

which, in this context, might represent a true state-evolution equation and an erroneous one, say one with the true parameter set and one with the erroneous set. Let $x_p(t, x)$ and $x_q(t, x)$ denote the solutions of these two equations, respectively, with initial state

$$x_p(0, x) = x_q(0, x) = x_0(x) \geq 0.$$ 

Then, the solutions of (12) for $i = p, q$, can be written as

$$x_i(t, x) = \int P_{ut, Dt}(x-y)x_0(y)dy, \text{ for } i = p, q.$$ 

In other words, each $x_i(t, x)$ is a convolution of a Gaussian kernel with $x_0(y)$.

A possible strategy to transport the mass from the distribution $x_p(t, \cdot)$ to the distribution $x_q(t, \cdot)$ is to transfer separately $P_{ut, Dt_p}(\cdot - y)$ to $P_{ut, Dt_q}(\cdot - y)$ for each $y$ and $t_p$, $t_q$.
average the individual costs with weight $x_0(y)dy$. This observation leads to the following inequality

$$
\mathcal{T}(x_p(t, \cdot), x_q(t, \cdot)) \leq \int \mathcal{T}(P_{up,t,D_p}t(\cdot - y), P_{up,t,D_p}t(\cdot - y))x_0(y)dy
$$

$$
= \int \mathcal{T}(P_{up,t,D_p}t(\cdot), P_{up,t,D_p}t(\cdot))x_0(y)dy
$$

since the OMT cost $\mathcal{T}(P_{up,t,D_p}t(\cdot), P_{up,t,D_p}t(\cdot))$ between the Gaussian distributions $P_{up,t,D_p}t(\cdot)$ and $P_{up,t,D_p}t(\cdot)$ is not affected when we simultaneously transport both by $y$. The OMT-metric between two independent Gaussian distributions has the following closed form expression

$$
\mathcal{T}(P_{up,t,D_p}t, P_{up,t,D_p}t) = (u_q - u_p)^2t^2 + (\sqrt{2D_q} - \sqrt{2D_p})^2t,
$$

[see e.g., Knott and Smith, 1984], and therefore (13) leads to

$$
\mathcal{T}(x_p(t, \cdot), x_q(t, \cdot)) \leq \|x_0\|_1 \left((u_q - u_p)^2t^2 + (\sqrt{2D_q} - \sqrt{2D_p})^2t\right),
$$

where $\|x_0\|_1$ is the total mass of the initial state, i.e. $\|x_0\|_1 = \int x_0(y)dy$, as $x_0(y) \geq 0$. Inequality (14) provides an upper bound for the transportation distance between the solutions of the advection–diffusion equation with different parameters. In particular it implies that, in a short time interval, small errors in the advection–diffusion parameters lead to small errors in the states in term of the OMT-metric. Notice that the upper bound (14) is a continuous function of the error in the model parameters $(u, \sqrt{D})$, which is tight when the initial state corresponds to a pulse (Dirac delta).

On the contrary, as previously explained (see Figure 1), we recall that the $\ell_2$-norm was unable to capture errors in the shift parameter, as it saturates to a certain constant value when the union of the support sets is an empty set.
5.2. 1-D advection-diffusion model

In this example, the proposed approach is tested by hypothesizing an erroneous constant advection parameter. It is also assumed that the measurements are downsampled versions of the states (observation operator $H$ as in (18) below).

Consider a one-dimensional state $x$ evolving according to (10). Let the spatial and time resolutions be $\delta_x = 0.025$ and $\delta_t = 1$, respectively and

$$x_k = \begin{bmatrix} x(x_1, k\delta_t) \\ \vdots \\ x(x_N, k\delta_t) \end{bmatrix}$$  \hspace{1cm} (15)

In addition, let us assume that the state is contaminated by additive Gaussian noise and thus (10) leads to the following difference equation for the state evolution

$$x_{k+1} = Fx_k + w_k,$$  \hspace{1cm} (16)

where, according to (11), the matrix $F$ has the $(i, j)$th entry given by

$$F(i, j) = \frac{\delta_x}{\sqrt{4\pi D}} e^{-\frac{(x_i - x_j - u)^2}{4D}}.$$  \hspace{1cm} (17)

In this example, $x$ is considered to have support on $[0, 5]$, hence, numerically, $x$ is a vector in $\mathbb{R}^{200 \times 1}$ and $F \in \mathbb{R}^{200 \times 200}$.

We assume that the true advection and diffusion process evolves with the following (but unknown) parameters

$$u(t) = 0.75, D(t) = 0.02, \forall t,$$

and the error $w_k$ can be well explained by a zero-mean Gaussian noise with covariance $Q = 0.2 I$. Note that we chose $t = 0$ as a reference point to denote the initial time of interest (e.g., $t_0 = 0$). Moreover, we assume that the measurement model is

$$y_k = Hx_k + v_k$$
with the following observation operator:

\[
H = \frac{1}{4} \begin{bmatrix}
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\
& & & & & & & \\
& & & & & & & \\
& & & & & & & \\
\end{bmatrix} \in \mathbb{R}^{50 \times 200}.
\] (18)

Selection of the above operator resembles a low-resolution sensor, which can only capture the mean of four neighboring elements of the state vector. Here, the observation error \(v_k\) is considered to be a zero–mean Gaussian distribution with covariance \(R = 0.2I\).

To make the problem formulation complete, let the initial state be given by

\[
x_0(x) = \begin{cases} 
10 & \text{for } 1.25 \leq x \leq 2.5 \\
0 & \text{otherwise}
\end{cases}
\]

which is shown by the solid line in the left panel of Figure 3. This initial state might be hypothesized as a pollutant concentration pulse propagated through a medium, a pulse of heat flux propagated through a soil column, or a convective mass of moisture evolving through the atmosphere.

Now assume that the model parameters are not exactly known. In particular, we used an erroneous (biased) advection-diffusion model with parameters

\[
\hat{u}(t) = 0.5, \hat{D}(t) = 0.02, \forall t.
\]

The background state \(x_b\), being a model prediction from a previous time step, is also biased and lags behind the true state by \(\hat{u}(t) - u(t) = 0.25\) as shown by the dotted line in the left panel of Figure 3. We consider that three measurements \(y_0, y_1, y_2\) are available (here obtained by adding zero-mean Gaussian noise with covariance \(R = 0.2I\) to the model states predicted with the correct model), and these are plotted in solid blue line, dash-dotted red line and dashed green line in the right panel of Figure 3.

We examined the estimation using the \(\ell_1\)-norm regularized 4D–Var formulation in (4) and the proposed formulation in (9). The estimated states of the two cost functions are...
demonstrated in Figure 4. The left panel of Figure 4 demonstrates the true and estimated state \( x_0 \). It is apparent that the procedure employing the OMT-metric is very effective in removing the bias and preserving the shape of the state. The estimated \( \hat{x}_0^{\text{omt+} \ell_1} \) is much closer to the true state than the estimate \( \hat{x}_0^{\ell_2+ \ell_1} \) given by the \( \ell_1 \)-norm regularized classical 4D–Var in (4). From the middle and the right panels, it is also apparent that the estimated states (at time step \( k = 1, 2 \)) \( \hat{x}_1^{\ell_2} \) and \( \hat{x}_2^{\ell_2} \) are more biased and more diffused than those of the \( \hat{x}_1^{\text{omt}} \) and \( \hat{x}_2^{\text{omt}} \) states given by the proposed method.

In order to provide a quantitative comparison, we computed the normalized squared error (NSE),

\[
\text{NSE} := \frac{\| \hat{x} - x \|^2}{\| x \|^2}
\]

recognizing its limitation in capturing the type of discrepancies that motivated us to introduce the OMT-formalism in the first place. The normalized-squared errors of the three states using the \( \ell_2 \)-based method (4) are NSE\(_{\ell_2, x_0} = 0.12 \), NSE\(_{\ell_2, x_1} = 0.12 \) and NSE\(_{\ell_2, x_2} = 0.20 \) respectively, while for the proposed method in equation (9), this error is NSE\(_{\text{omt}, x_0} = 0.03 \), NSE\(_{\text{omt}, x_1} = 0.02 \) and NSE\(_{\text{omt}, x_2} = 0.05 \), respectively. It is worth noting that the computation of this example was implemented using \textit{cvx} optimization toolbox in Matlab \([\text{Grant et al., 2012}]\) on a desktop with a 2.4GHz CPU clock rate. It took 9.7 seconds to get the result for the \( \ell_2 \)-based method while the OMT-based method required more than 88 seconds, highlighting the need to develop more computationally efficient algorithms for application to real problems.

In the above example, the \( \ell_1 \)-norm regularization evoked for the estimation of the initial state \( x_0 \) which exhibits sparsity, was implemented using the following linear transformation...
Φ as stated in problem (9):

\[
\Phi = \begin{bmatrix}
1 & -1 & 0 & 0 & \ldots & 0 \\
0 & 0 & 1 & -1 & 0 & \ldots \\
\vdots & \vdots & \vdots & \vdots & \ddots & \ddots \\
\end{bmatrix}.
\]

As is evident, the above choice of the Φ transformation is a first order differencing operator. This particular choice of Φ refers back to our prior knowledge with respect to the degree of smoothness and sparsity of the first order derivative of \(x_0\). In other words, one can see that the state of interest \(x_0\) is piece-wise constant and thus its first order differences form a sparse vector with a large number of zero elements. Indeed, the incorporation of the \(\ell_1\)-norm regularization of \(\|\Phi x_0\|_1\) is a reflection of this prior knowledge in our data assimilation scheme which leads to an effective removal of the high-frequency random errors and recovery of sharp jumps and singularities.

### 5.3. 2–D advection–diffusion model

In this example, the evolution of the state is dictated by a two–dimensional advection–diffusion equation

\[
\partial_t x + u_x(t) \partial_x x + u_y(t) \partial_y x = D_x \partial_{xx} x + D_y \partial_{yy} x \tag{19}
\]

where \(u_x, u_y\) and \(D_x, D_y\) represent the advection and diffusion parameters in the \(x\) and \(y\) direction, respectively. We discretize (19) with time and spatial resolutions \(\delta_t = 1\) and \(\delta_x = \delta_y = 0.1\), respectively. If we define

\[
X_k = \begin{bmatrix}
x(x_1, y_1, k\delta_t) & \ldots & x(x_1, y_N, k\delta_t) \\
x(x_N, y_1, k\delta_t) & \ldots & x(x_N, y_N, k\delta_t)
\end{bmatrix}
\]

then (19) leads to the following difference equation for the evolution of the state

\[
X_{k+1} = F_{x,k} X_k F_{y,k}^T + W_k
\]
where $F_{x,k}$ and $F_{y,k}$ are given by (17) with advection and diffusion parameters set to $u_x, D_x$ and $u_y, D_y$, respectively. In this example, we consider an advection-diffusion model with parameters

$$u_x(t) = u_y(t) = 0.15 \text{ and } D_x(t) = D_y(t) = 0.02, \forall t.$$  

and the process noise is considered zero mean Gaussian with covariance $Q = 0.1I$.

The measurement equation is cast as follows:

$$Y_k = HX_kH^T + V_k$$  

(21)

where $H$ is as in (18) but in a two dimensional setting, i.e., the sensor output is a noisy and averaged representation of the true state over a box of size $2 \times 2$. In this example, $F_{x,k}, F_{y,k} \in \mathbb{R}^{12 \times 12}$ and $H \in \mathbb{R}^{6 \times 12}$ and the observation noise is also zero mean Gaussian with covariance $R = 0.1I$.

The true state $X_0$ is chosen to be piecewise constant as shown in the top left most panel of Figure 6 and the subsequent states $X_1$ and $X_2$ are generated with the true model. Once again, due to some imperfect knowledge of the model, we are led to consider the following erroneous advection and diffusion parameters

$$\tilde{u}_x(t) = \tilde{u}_y(t) = 0.1 \text{ and } \tilde{D}_x(t) = \tilde{D}_y(t) = 0.015, \forall t,$$

(a biased model) and seek to estimate the true states $X_0, X_1$ and $X_2$. The background state $X_b$, being itself a model prediction from the previous time step, is also biased (it lags behind and is less diffused than the true state $X_0$ according to the biased model as shown in the left panel of Figure 5). Using the background state $X_b$ and the available measurements up to time $t = 2$, shown in Figure 5, we estimate the states $X_0, X_1, X_2$ by solving the weak-constraint 4D-Var problem and the variational problem based on the
OMT–metric. In both cases we also use suitable regularizations to promote the shape and the smoothness of the states.

The estimation results are shown in Figure 6 along with the true states at time \( t = 0, 1, 2 \). Once again, it can be seen that the proposed methodology produces improved estimates of the true states. This fact seems more apparent in the estimation quality of the initial state in the first row panels of Figure 6, in which not only the bias but also the involved random error are well removed and the true shape is properly recovered. It is apparent that the OMT metric captures more accurately the drift and diffusion and facilitates extracting more accurate estimates from subsequent observations. The normalized squared errors of the estimated states using (4) are respectively \( \text{NSE}_{\ell_2, X_0} = 0.20 \), \( \text{NSE}_{\ell_2, X_1} = 0.06 \) and \( \text{NSE}_{\ell_2, X_2} = 0.10 \) while the proposed method in equation (9) leads to \( \text{NSE}_{\text{omt}, X_0} = 0.03 \), \( \text{NSE}_{\text{omt}, X_1} = 0.04 \) and \( \text{NSE}_{\text{omt}, X_2} = 0.09 \). The computation time for the \( \ell_2 \)-based method and the OMT-based method are 8.5 seconds and 43 seconds, respectively.

6. Conclusions

Environmental models often suffer from structural errors due to inadequate and/or over simplified characterization of the underlying physics. In this paper, we presented an approach which allows us to take into account non-parametrically both random and in particular systematic model error in data assimilation using an unbalanced Optimal Mass Transport (OMT) metric. Using 1-D and 2-D advection-diffusion dynamics, we specifically elaborated on the effectiveness of the OMT-metric for tackling systematic model error. We also showed that the OMT-metric in combination with \( \ell_1 \)-regularization can be very useful for estimation of state variables exhibiting singular structures with a sparse representation in a properly chosen transform domain, e.g., in a derivative domain.
The formalism proposed in here can be considered as an extension to the classic variational data assimilation that can allow us to non-parametrically account for the model systematic error in data assimilation. We suggest that the proposed OMT-equipped variational data assimilation formalism has the potential to handle problems with complicated model error dynamics and deserves further study both in theory and in its application to geophysical data assimilation problems.

Acknowledgments. This work has been partially supported by NASA (GPM award NNX10AO12G, and an Earth and Space Science Fellowship-NNX12AN45H to the third author), NSF (Grant ECCS 1027696), AFOSR (Grant FA9550-12-1-0319), and the Ling and Vincentine Hermes-Luh Endowments.

References


Benamou, J., and Y. Brenier (2001), Mixed $L_2$-Wasserstein optimal mapping between
prescribed density functions, Journal of Optimization Theory and Applications, 111(2),
255–271.

from reanalysis data?, Journal of Geophysical Research: Atmospheres, 109(D11), doi:
10.1029/2004JD004536.

Bouttier, F., and P. Courtier (1999), Data assimilation concepts and methods, Meteorolo-
logical training course lecture series. ECMWF.

Bouttier, F., and P. Courtier (2002), Data assimilation concepts and methods, Meteorolo-
logical training course lecture series. ECMWF.

Caparrini, F., F. Castelli, and D. Entekhabi (2004), Estimation of surface turbulent fluxes
through assimilation of radiometric surface temperature sequences, Journal of Hydrom-
eteorology, 5(1), 145–159.

Carrassi, A., and S. Vannitsem (2010), Accounting for model error in variational data

Courtier, P., and O. Talagrand (1990), Variational assimilation of meteorological obser-
vations with the direct and adjoint shallow-water equations, Tellus A, 42(5), 531–549.

Crow, W. T., and E. F. Wood (2003), The assimilation of remotely sensed soil brightness
temperature imagery into a land surface model using ensemble kalman filtering: A case
study based on estar measurements during sgp97, Advances in Water Resources, 26(2),
137–149.


Donoho, D. (2006b), For most large underdetermined systems of linear equations the minimal $\ell_1$-norm solution is also the sparsest solution, *Communications on pure and applied mathematics, 59*(6), 797–829.


Hansen, P. C. (2010), *Discrete inverse problems: insight and algorithms*, vol. 7, SIAM.


Liu, Y., A. Weerts, M. Clark, H.-J. Hendricks Franssen, S. Kumar, H. Moradkhani, D.-J.
Seo, D. Schwanenberg, P. Smith, A. van Dijk, et al. (2012), Advancing data assimilation
in operational hydrologic forecasting: progresses, challenges, and emerging opportuni-
ties, *Hydrology and Earth System Sciences*, 16(10), 3863–3887.

Lorenc, A. (1981), A global three-dimensional multivariate statistical interpolation


Lorenc, A. C. (2003), The potential of the ensemble Kalman filter for NWP – a comparison
with 4D–Var, *Quarterly Journal of the Royal Meteorological Society*, 129(595), 3183–
3203.

Lorenc, A. C., S. P. Ballard, R. S. Bell, N. B. Ingleby, P. L. F. Andrews, D. M. Barker,
J. R. Bray, A. M. Clayton, T. Dalby, D. Li, T. J. Payne, and F. W. Saunders (2000),
The Met. Office global three-dimensional variational data assimilation scheme, *Quart.

Martin, M. J., M. J. Bell, and N. K. Nichols (2002), Estimation of systematic error in
an equatorial ocean model using data assimilation, *International Journal for Numerical

McLaughlin, D. (2002), An integrated approach to hydrologic data assimilation: interpo-

Monge, G. (1781), *Mémoire sur la théorie des déblais et des remblais*, De l’Imprimerie
Royale.


Vidard, P., A. Piacentini, and F.-X. LE DIMET (2004), Variational data analysis with control of the forecast bias, Tellus A, 56(3), 177–188.


Figure 1. Left panel: initial state $x_0$, true $x_1$, and incorrect states $\tilde{x}_{1}^{(i)}$, $i \in \{1, 2, 3, 4\}$ obtained by an erroneous model, here, a biased estimate of the shift coefficient at time $t = 1$; Right panel: quadratic norm $\|x_1 - \tilde{x}_{1}^{(i)}\|^2_2$ and OMT-metric $T(x_1, \tilde{x}_{1}^{(i)})$ between the true and incorrect states at time $t = 1$ for the four erroneous values of the shift coefficient shown in the left panel. Observe how the quadratic metric is insensitive to the model error when the supports of the true and erroneous states are not overlapping while the OMT-metric increases monotonically proportionally to the model error (here a shift).
Figure 2. Left panel: true density $x$ (in the solid line) and two oppositely biased and noisy densities $x_b$ (in the dashed line) and $y$ (in the dash-dotted line). Right panel: true density $x$ (in the solid blue line) and the $\ell_2$ ($\hat{x}^{\ell_2}$ in the dashed green line) and OMT-estimates ($\hat{x}^{\text{omt}}$ in dash-dotted green line), respectively.

Figure 3. Left panel: initial state $x_0$ (solid line), background state $x_b$ (dotted line), and two model simulated states $x_1, x_2$ at time steps $t = 1, 2$ (dash-dotted and dashed lines, respectively). The simulated states $x_1, x_2$ were produced from $x_0$ using the advection diffusion model with parameters $u = 0.75$ and $D = 0.02$, while $x_b$ was taken to be a lagged and diffused variant of $x_0$. Right panel: noisy observations $y_0, y_1, y_2$ at time steps $t = 0, 1, 2$ produced from $x_0, x_1, x_2$, respectively. The components of the process and observation noises were assumed Gaussian with zero mean and variance $R = 0.2$.

Figure 4. From left to right: true states $x_t$ (solid blue line) and estimated states $\hat{x}_t$, at time steps $t = 0, 1, 2$ in successive panels, using $\ell_2$ and OMT and the erroneous advection diffusion model with parameters $\tilde{u} = 0.5$ and $\tilde{D} = 0.02$. At $t = 0$, $\ell_1$ regularization has also been used to promote the sparseness of the state in the derivative space. The accuracy of the OMT estimates is notable.

Figure 5. From left to right: the biased background state $X_b$ (produced as a prediction from a previous time step using the biased model $\tilde{u}_x = \tilde{u}_y = 0.1$ and $\tilde{D}_x = \tilde{D}_y = 0.015$) and the observations $Y_0, Y_1$ and $Y_2$ at time steps $t = 0, 1, 2$. The observations $Y_k = HX_kH^T + V_k$ are downsampled and noisy version of the states $X_k$ produced by the true model with parameters $u_x = u_y = 0.15$, and $D_x = D_y = 0.02$, while the entries of the observation noise $V_k$ are zero-mean Gaussian random variables with variance 0.1.
Figure 6. First row panels, from left to right: true state $X_0$ and its estimates via weak-constraint 4D-Var plus $\ell_1$-norm regularization, and via the proposed OMT–based methodology, based on the biased background state $X_b$ and the observations $Y_k$ (shown in Figure 5). Note that in all cases the biased model with parameters $\tilde{u}_x = \tilde{u}_y = 0.1$ and $\tilde{D}_x = \tilde{D}_y = 0.015$ was used for the estimation. The second and third row panels, from left to right, show the true and estimated states at time steps $t = 1$ and $t = 2$. Observe how OMT–based 4D–Var accurately recovers the true state, especially at time $t = 0$, clearly outperforming the weak–constraint 4D–Var.
The image contains a graph with the following annotations:

- On the left side, there are vertical bars labeled $x_0$, $x_1$, and $\tilde{x}_1$, with a bar labeled $\tilde{x}_1^{(1)}$.
- On the right side, there is a graph with a plotted line showing $\|x_1 - \tilde{x}_1^{(i)}\|^2$ and $\mathcal{T}(x_1 - \tilde{x}_1^{(i)})$.

The graph illustrates a comparison between different values and their transformations.
$t = 0$

$X_0$

$\hat{X}_0^{\ell_2 + \ell_1}$

$\hat{X}_0^{\text{omt} + \ell_1}$

$X_1$

$\hat{X}_1^{\ell_2}$

$\hat{X}_1^{\text{omt}}$

$t = 1$

$X_2$

$\hat{X}_2^{\ell_2}$

$\hat{X}_2^{\text{omt}}$

$t = 2$