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]

[1]Department of Electrical and Computer Engineering, University of Minnesota, 200 Union Street SE, Minneapolis, MN, USA

[2]Department of Civil Engineering, St. Anthony Falls Laboratory, and National Center for Earth-Surface Dynamics, University of Minnesota, 2 Third Avenue SE, Minneapolis, MN, USA

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as an 'Accepted Article', doi: 10.1002/2013WR014966

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 ℓ_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:

$$\mathbf{x}_{k+1} = F_k(\mathbf{x}_k) + \mathbf{w}_k \tag{1a}$$

$$\mathbf{y}_k = H_k(\mathbf{x}_k) + \mathbf{v}_k, \tag{1b}$$

where $F(\cdot), H(\cdot)$ denote the model and observation operators, respectively, and the (column) vectors $\mathbf{w}, \mathbf{v}$ represent process and observation noise. A prior estimate of the true initial state $\mathbf{x}_0$ in data assimilation is typically referred to as the "background state" $\mathbf{x}_b$, 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 $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_K$. In the statistical filtering theory, this problem is referred to as a *fixed-interval smoothing problem*. If the error components $\mathbf{w}_k, \mathbf{v}_k$ ($k \in \{0, \ldots, K\}$) and $\mathbf{x}_0 - \mathbf{x}_b$ 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 $\mathbf{w}$, estimation of the underlying states amounts to minimizing the following

quadratic expression:

$$\min_{\mathbf{x}_0} \left\{ \|\mathbf{x}_0 - \mathbf{x}_b\|^2_{P^{-1}} + \sum_{k=0}^{K} \|\mathbf{y}_k - H_k F_k \dots F_1 F_0(\mathbf{x}_0)\|^2_{R_k^{-1}} \right\}, \tag{2}$$

where P and R_k represent the covariances of the background error $\mathbf{e} := \mathbf{x}_0 - \mathbf{x}_b$ and $\mathbf{v}_k$, respectively. Here, the notation $\|\mathbf{e}\|^2_{P^{-1}}$ represents the weighted quadratic norm, that is $\mathbf{e}^T P^{-1} \mathbf{e}$, 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 $\mathbf{w}$ is incorporated which leads to the following problem formulation, which is often called *weak constraint* 4D-Var:

$$\min_{\mathbf{x}_0,\dots,\mathbf{x}_K} \left\{ \|\mathbf{x}_0 - \mathbf{x}_b\|^2_{P^{-1}} + \sum_{k=0}^{K} \|\mathbf{y}_k - H_k(\mathbf{x}_k)\|^2_{R_k^{-1}} + \sum_{k=1}^{K} \|\mathbf{x}_k - F_{k-1}(\mathbf{x}_{k-1})\|^2_{Q_k^{-1}} \right\}, \tag{3}$$

where similarly Q_k represents the covariance of $\mathbf{w}_k$.

More recently, the use of ℓ_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 $\mathbf{x}_0$ is "smooth," except for a few distinct jumps or that the derivative of $\mathbf{x}_0$ has a Laplace-like distribution, then the ℓ_1–norm $\|\Phi(\mathbf{x}_0)\|_1$ with Φ a linear derivative-like operator (e.g., wavelet transform) can be used as a regularization term as follows:

$$\min_{\mathbf{x}_0,\dots,\mathbf{x}_K} \left\{ \|\mathbf{x}_0 - \mathbf{x}_b\|^2_{P^{-1}} + \sum_{k=0}^{K} \|\mathbf{y}_k - H_k(\mathbf{x}_k)\|^2_{R_k^{-1}} + \sum_{k=1}^{K} \|\mathbf{x}_k - F_{k-1}(\mathbf{x}_{k-1})\|^2_{Q_k^{-1}} + \gamma \|\Phi(\mathbf{x}_0)\|_1 \right\}, \tag{4}$$

where γ 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 $\mathbf{w}$. 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. $\sum_{i=1}^{N} \mathbf{x}(x_i) = \sum_{i=1}^{N} \tilde{\mathbf{x}}(\tilde{x}_i)$, 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 $\mathbf{x}$ to the distribution $\tilde{\mathbf{x}}$. 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 $\mathcal{T}(\mathbf{x}, \tilde{\mathbf{x}})$ can be computed as follows:

$$\begin{aligned} \mathcal{T}(\mathbf{x}, \tilde{\mathbf{x}}) \;:=\; & \min_{m} \sum_{i,j} c_{i,j}\, m(x_i, \tilde{x}_j) \\ \text{subject to} \quad & m(x_i, \tilde{x}_j) \geq 0, \\ & \sum_{j} m(x_i, \tilde{x}_j) = \mathbf{x}(x_i), \\ & \sum_{i} m(x_i, \tilde{x}_j) = \tilde{\mathbf{x}}(\tilde{x}_j). \end{aligned} \tag{5}$$

The above formulation can be expressed more compactly if we let $\mathbf{1}$ 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,

$$\begin{aligned} \mathcal{T}(\mathbf{x}, \tilde{\mathbf{x}}) \;:=\; & \min_{M} \operatorname{tr}(CM) \\ \text{subject to} \quad & M\mathbf{1} = \mathbf{x},\; M^T\mathbf{1} = \tilde{\mathbf{x}}, M(i,j) \geq 0 \;\forall i,j. \end{aligned} \tag{6}$$

Here, the matrix M is a joint density matrix with positive elements, often referred to as the *transportation plan* [*Villani*, 2003], where $\mathbf{x}$ and $\tilde{\mathbf{x}}$ 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 ℓ_2-norm as follows:

$$\mathcal{T}_\sigma(\mathbf{x},\tilde{\mathbf{x}}) := \min_{\hat{\mathbf{x}}} \left\{ \mathcal{T}(\hat{\mathbf{x}},\tilde{\mathbf{x}}) + \sigma \|\mathbf{x}-\hat{\mathbf{x}}\|_2^2 \right\}, \tag{7}$$

where the non-negative parameter σ is used to denote the relative significance of the ℓ_2-cost. Notice that the above relaxed OMT metric is obtained through a nested minimization problem based on an intermediate state $\hat{\mathbf{x}}$ with minimal ℓ_2 distance to the true state $\mathbf{x}$ (second term within the bracket), while at the same time its mass is considered to be equal to $\tilde{\mathbf{x}}$, through the use of the classic $\mathcal{T}(\hat{\mathbf{x}},\tilde{\mathbf{x}})$ metric (first term within the bracket). In other words, this new relaxed OMT metric penalizes the distance of $\hat{\mathbf{x}}$ from the model predicted state $\tilde{\mathbf{x}}$ in the OMT sense, while it keeps the solution close enough to the true state $\mathbf{x}$ without strictly enforcing mass equality due to the presence of model error. Obviously, the parameter σ 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 $\mathcal{T}(\mathbf{x},\tilde{\mathbf{x}})$ and $\|\mathbf{x}-\tilde{\mathbf{x}}\|_2$ 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 (oppositively) 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 $\mathbf{x}_k(x)$ at time k represents a density function with a compact support in a bounded set $\mathcal{D}$ where $\mathbf{x}_k(x) > 0$ for $\forall x \in \mathcal{D}$. To this end, consider a simple non-diffusive advective state space as follows:

$$\mathbf{x}_{k+1}(x) = \mathbf{x}_k(x - d),$$

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

$$\mathbf{x}_0 = \begin{cases} 10 & \text{for} -0.5 \leq x \leq 0.5 \\ 0 & \text{otherwise,} \end{cases}$$

which is shown in the black solid line in Figure 1, while the state $\mathbf{x}_1(x)$ 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 $\left\{\tilde{\mathbf{x}}_1^{(i)}\right\}_{i=1}^{4}$ 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 $\mathbf{x}_1$ and $\tilde{\mathbf{x}}_1$ are not overlapping, $\|\mathbf{x}_1 - \tilde{\mathbf{x}}_1\|_2^2$ 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

ℓ_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 $\mathbf{x}$ and two biased and noisy estimates on either side. The right panel displays the reconciled estimate of $\mathbf{x}$ obtained via averaging the two biased states using the ℓ_2 norm, $\hat{\mathbf{x}}^{\ell_2}$, and using an OMT metric, $\hat{\mathbf{x}}^{\text{omt}}$, and compares these with the true density $\mathbf{x}$. In essence, the reconciled estimate represents a "mean" (in the ℓ_2-sense and in the OMT-sense, respectively). Naturally, the ℓ_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 $\mathbf{x}_k$ and the predicted state $F_{k-1}(\mathbf{x}_{k-1})$ as $\mathcal{T}(\mathbf{x}_k, F_{k-1}(\mathbf{x}_{k-1}))$ and, likewise, the transportation distance between the true and background states as $\mathcal{T}(\mathbf{x}_0, \mathbf{x}_b)$. Our proposed formulation of the variational data assimilation problem is as follows:

$$\min_{\mathbf{x}_0,\ldots,\mathbf{x}_K} \left\{ \mathcal{T}(\mathbf{x}_0, \mathbf{x}_b) + \sum_{k=0}^{K} \|\mathbf{y}_k - H_k(\mathbf{x}_k)\|^2_{R_k^{-1}} + \sum_{k=1}^{K} \mathcal{T}(\mathbf{x}_k, F_{k-1}(\mathbf{x}_{k-1})) \right\} \tag{8}$$

where the ℓ_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 $\mathcal{T}$'s by $\mathcal{T}_\sigma$'s to allow for the possibility of non-conservative states, is as follows:

$$\min_{\mathbf{x}_0,\ldots,\mathbf{x}_K}\left\{\mathcal{T}_\sigma(\mathbf{x}_0,\mathbf{x}_b)+\sum_{k=0}^{K}\|\mathbf{y}_k-H_k\mathbf{x}_k\|^2_{R_k^{-1}}+\sum_{k=1}^{K}\mathcal{T}_\sigma(\mathbf{x}_k,F_{k-1}\mathbf{x}_{k-1})+\gamma\|\Phi\mathbf{x}_0\|_1\right\}.$$

This includes also a regularization term $\|\Phi\mathbf{x}_0\|_1$ 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

$$\begin{aligned}\min_{\substack{\mathbf{x}_0,\ldots,\mathbf{x}_K\\ \tilde{\mathbf{x}}_b,\tilde{\mathbf{x}}_1,\ldots,\tilde{\mathbf{x}}_K}}\Bigg\{&\sigma\left(\|\mathbf{x}_0-\tilde{\mathbf{x}}_b\|^2+\sum_{k=1}^{K}\|\mathbf{x}_k-\tilde{\mathbf{x}}_k\|^2\right)+\sum_{k=0}^{K}\|\mathbf{y}_k-H_k\mathbf{x}_k\|^2_{R_k^{-1}}+\gamma\|\Phi\mathbf{x}_0\|_1\\&+\mathcal{T}(\tilde{\mathbf{x}}_b,\mathbf{x}_b)+\sum_{k=1}^{K}\mathcal{T}(\tilde{\mathbf{x}}_k,F_{k-1}\mathbf{x}_{k-1})\Bigg\}\end{aligned}\tag{9}$$

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

$$\begin{aligned}\mathcal{T}(\tilde{\mathbf{x}}_k,F_{k-1}\mathbf{x}_{k-1}) \;=\; &\min_{M_k}\operatorname{tr}(CM_k)\\ &\text{subject to}\qquad M_k(i,j)\geq 0,\ M_k\mathbf{1}=\hat{\mathbf{x}}_k,\ M_k^T\mathbf{1}=F_{k-1}\mathbf{x}_{k-1}.\end{aligned}$$

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

$$\begin{aligned}\min_{\substack{\mathbf{x}_0,\ldots,\mathbf{x}_K\\ M_b,M_1,\ldots,M_K}}\Bigg\{&\sigma\left(\|\mathbf{x}_0-M_b\mathbf{1}\|^2+\sum_{k=1}^{K}\|\mathbf{x}_k-M_k\mathbf{1}\|^2\right)+\sum_{k=0}^{K}\|\mathbf{y}_k-H_k\mathbf{x}_k\|^2_{R_k^{-1}}+\\&\operatorname{tr}\left(C(M_b+\sum_{k=1}^{K}M_k)\right)+\gamma\|\Phi\mathbf{x}_0\|_1\Bigg\}\\ \text{subject to}\quad &M_b(i,j),M_k(i,j)\geq 0,\ M_b^T\mathbf{1}=\mathbf{x}_0,\ M_k^T\mathbf{1}=F_{k-1}\mathbf{x}_{k-1},\ k=1,\ldots,K,\end{aligned}$$

where we replaced $\hat{\mathbf{x}}_b$ by $M_b\mathbf{1}$ and $\tilde{\mathbf{x}}_k$ by $M_k\mathbf{1}$. We note that the size of M_k's is $N \times N$ since $\mathbf{x}_k$'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

$$\partial_t \mathbf{x} + u\,\partial_x \mathbf{x} = D\partial_{xx}\mathbf{x} \tag{10}$$

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

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

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 $\mathbf{x}_0$, the advection–diffusion equation (10) is linear in $\mathbf{x}$, with the following solution:

$$\mathbf{x}(t,x) = \int P_{ut,Dt}(x-y)\mathbf{x}_0(y)dy\,. \tag{11}$$

Now, let us consider the advection–diffusion equations

$$\partial_t \mathbf{x} + u_i\,\partial_x \mathbf{x} = D_i\,\partial_{xx}\mathbf{x},\ \text{for } i = p,\, q\,, \tag{12}$$

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 $\mathbf{x}_p(t,x)$ and $\mathbf{x}_q(t,x)$ denote the solutions of these two equations, respectively, with initial state

$$\mathbf{x}_p(0,x) = \mathbf{x}_q(0,x) = \mathbf{x}_0(x) \geq 0.$$

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

$$\mathbf{x}_i(t,x) = \int P_{u_i t, D_i t}(x-y)\mathbf{x}_0(y)dy, \quad \text{for } i = p,\, q\,.$$

In other words, each $\mathbf{x}_i(t,x)$ is a convolution of a Gaussian kernel with $\mathbf{x}_0(y)$.

A possible strategy to transport the mass from the distribution $\mathbf{x}_p(t,\cdot)$ to the distribution $\mathbf{x}_q(t,\cdot)$ is to transfer separately $P_{u_p t, D_p t}(\cdot - y)$ to $P_{u_q t, D_q t}(\cdot - y)$ for each y and

average the individual costs with weight $\mathbf{x}_0(y)dy$. This observation leads to the following inequality

$$\begin{aligned}\mathcal{T}(\mathbf{x}_p(t,\cdot),\mathbf{x}_q(t,\cdot)) &\leq \int \mathcal{T}(P_{u_pt,D_pt}(\cdot-y),P_{u_qt,D_qt}(\cdot-y))\mathbf{x}_0(y)dy \\ &= \int \mathcal{T}(P_{u_pt,D_pt}(\cdot),P_{u_qt,D_qt}(\cdot))\mathbf{x}_0(y)dy \\ &= \mathcal{T}(P_{u_pt,D_pt}(\cdot),P_{u_qt,D_qt}(\cdot)) \int \mathbf{x}_0(y)dy\end{aligned} \tag{13}$$

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

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

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

$$\mathcal{T}(\mathbf{x}_p(t,\cdot),\mathbf{x}_q(t,\cdot)) \leq \|\mathbf{x}_0\|_1 \left((u_q-u_p)^2t^2 + (\sqrt{2D_q}-\sqrt{2D_p})^2t\right), \tag{14}$$

where $\|\mathbf{x}_0\|_1$ is the total mass of the initial state, i.e. $\|\mathbf{x}_0\|_1 = \int \mathbf{x}_0(y)dy$, as $\mathbf{x}_0(y) \geq 0$. 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 ℓ_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 $\mathbf{x}$ evolving according to (10). Let the spatial and time resolutions be $\delta_x = 0.025$ and $\delta_t = 1$, respectively and

$$\mathbf{x}_k = \begin{bmatrix} \mathbf{x}(x_1, k\delta_t) \\ \vdots \\ \mathbf{x}(x_N, k\delta_t). \end{bmatrix} \tag{15}$$

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

$$\mathbf{x}_{k+1} = F\mathbf{x}_k + \mathbf{w}_k, \tag{16}$$

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

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

In this example, $\mathbf{x}$ is considered to have support on $[0, 5]$, hence, numerically, $\mathbf{x}$ 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 $\mathbf{w}_k$ 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

$$\mathbf{y}_k = H\mathbf{x}_k + \mathbf{v}_k$$

with the following observation operator:

$$H = \frac{1}{4}\begin{bmatrix} 1\,1\,1\,1\;0\,0\,0\,0\;\ldots\;0\,0\,0\,0 \\ 0\,0\,0\,0\;1\,1\,1\,1\;\ldots\;0\,0\,0\,0 \\ \vdots \qquad \vdots \qquad \vdots \qquad \vdots \end{bmatrix} \in \mathbb{R}^{50\times 200}. \tag{18}$$

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 $\mathbf{v}_k$ 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

$$\mathbf{x}_0(x) = \begin{cases} 10 & \text{for } 1.25 \leq x \leq 2.5 \\ 0 & \text{otherwise} \end{cases}$$

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

$$\tilde{u}(t) = 0.5, \tilde{D}(t) = 0.02, \forall t.$$

The background state $\mathbf{x}_b$, being a model prediction from a previous time step, is also biased and lags behind the true state by $\tilde{u}(t) - u(t) = 0.25$ as shown by the dotted line in the left panel of Figure 3. We consider that three measurements $\mathbf{y}_0, \mathbf{y}_1, \mathbf{y}_2$ 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 ℓ_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 $\mathbf{x}_0$. 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 $\hat{\mathbf{x}}_0^{\text{omt}+\ell_1}$ is much closer to the true state than the estimate $\hat{\mathbf{x}}_0^{\ell_2+\ell_1}$ given by the ℓ_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 $k = 1, 2$) $\hat{\mathbf{x}}_1^{\ell_2}$ and $\hat{\mathbf{x}}_2^{\ell_2}$ are more biased and more diffused than those of the $\hat{\mathbf{x}}_1^{\text{omt}}$ and $\hat{\mathbf{x}}_2^{\text{omt}}$ states given by the proposed method.

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

$$\text{NSE} := \frac{\|\hat{\mathbf{x}} - \mathbf{x}\|^2}{\|\mathbf{x}\|^2}$$

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 ℓ_2-based method (4) are $\text{NSE}_{\ell_2,\mathbf{x}_0} = 0.12$, $\text{NSE}_{\ell_2,\mathbf{x}_1} = 0.12$ and $\text{NSE}_{\ell_2,\mathbf{x}_2} = 0.20$ respectively, while for the proposed method in equation (9), this error is $\text{NSE}_{\text{omt},\mathbf{x}_0} = 0.03$, $\text{NSE}_{\text{omt},\mathbf{x}_1} = 0.02$ and $\text{NSE}_{\text{omt},\mathbf{x}_2} = 0.05$, respectively. It is worth noting that the computation of this example was implemented using *cvx* optimization toolbox in Matlab ® [*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 ℓ_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 ℓ_1-norm regularization evoked for the estimation of the initial state $\mathbf{x}_0$ which exhibits sparsity, was implemented using the following linear transformation

Φ as stated in problem (9):

$$\Phi = \begin{bmatrix} 1 & -1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & -1 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}.$$

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 $\mathbf{x}_0$. In other words, one can see that the state of interest $\mathbf{x}_0$ 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 ℓ_1-norm regularization of $\|\Phi\mathbf{x}_0\|_1$ 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

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

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

$$X_k = \begin{bmatrix} \mathbf{x}(x_1, y_1, k\delta_t) & \dots & \mathbf{x}(x_1, y_N, k\delta_t) \\ & & \\ \mathbf{x}(x_N, y_1, k\delta_t) & \dots & \mathbf{x}(x_N, y_N, k\delta_t) \end{bmatrix} \tag{20}$$

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:

$$Y_k = HX_kH^T + V_k \tag{21}$$

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 2×2. In this example, $F_{x,k}, F_{y,k} \in \mathbb{R}^{12\times12}$ and $H \in \mathbb{R}^{6\times12}$ 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 ℓ_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 ℓ_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

Anderson, B., and J. Moore (2012), *Optimal filtering*, Courier Dover Publications.

Auroux, D., and J. Fehrenbach (2011), Identification of velocity fields for geophysical fluids from a sequence of images, *Experiments in Fluids*, *50*(2), 313–328.

Bateni, S., and D. Entekhabi (2012), Surface heat flux estimation with the ensemble Kalman smoother: Joint estimation of state and parameters, *Water Resources Research*, *48*(8).

Benamou, J. (2003), Numerical resolution of an unbalanced mass transport problem, *ESAIM: Mathematical Modelling and Numerical Analysis*, *37*(05), 851–868.

Benamou, J., and Y. Brenier (2000), A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem, *Numerische Mathematik*, *84*(3), 375–393.

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.

Bengtsson, L., S. Hagemann, and K. I. Hodges (2004), Can climate trends be calculated 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, *Meteorological training course lecture series. ECMWF.*

Bouttier, F., and P. Courtier (2002), Data assimilation concepts and methods, *Meteorological 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 Hydrometeorology*, *5*(1), 145–159.

Carrassi, A., and S. Vannitsem (2010), Accounting for model error in variational data assimilation: A deterministic formulation, *Monthly Weather Review*, *138*(9), 3369–3386.

Courtier, P., and O. Talagrand (1990), Variational assimilation of meteorological observations 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.

Daley, R. (1993), *Atmospheric data analysis*, vol. 2, Cambridge university press.

Dee, D. P. (2005), Bias and data assimilation, *Quarterly Journal of the Royal Meteorological Society*, *131*(613), 3323–3343.

Derber, J., and A. Rosati (1989), A global oceanic data assimilation system, *Journal of Physical Oceanography*, *19*(9), 1333–1347.

Derber, J. C. (1989), A Variational Continuous Assimilation Technique, *Monthly Weather Review*, *117*(11), 2437–2446.

Donoho, D. (2006a), Compressed sensing, *IEEE Transactions on Information Theory*, *52*(4), 1289–1306.

Donoho, D. (2006b), For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution, *Communications on pure and applied mathematics*, *59*(6), 797–829.

Durand, M., and S. Margulis (2007), Correcting first-order errors in snow water equivalent estimates using a multifrequency, multiscale radiometric data assimilation scheme, *Journal of Geophysical Research: Atmospheres*, *112*(D13).

Ebtehaj, A., M. Zupanski, G. Lerman, and E. Foufoula-Georgiou (2014), Variational data assimilation via sparse regularization, *Tellus A*, *66*, 21789, doi: http://dx.doi.org/10.3402/tellusa.v66.21789.

Ebtehaj, A. M., and E. Foufoula-Georgiou (2013), On variational downscaling, fusion, and assimilation of hydrometeorological states: A unified framework via regularization, *Water Resources Research*, *49*, 1–20, doi:10.1002/wrcr.20424.

Fisher, M., M. Leutbecher, and G. Kelly (2005), On the equivalence between Kalman smoothing and weak-constraint four-dimensional variational data assimilation, *Quarterly Journal of the Royal Meteorological Society*, *131*(613), 3235–3246.

Fisher, R. A. (1922), On the mathematical foundations of theoretical statistics, *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, pp. 309–368.

Foufoula-Georgiou, E., A. M. Ebtehaj, S. Q. Zhang, and A. Y. Hou (2013), Downscaling satellite precipitation with emphasis on extremes: A Variational ℓ_1–norm Regularization in the Derivative Domain, *Surv. in Geophysics (to appear)*.

Freitag, M., N. Nichols, and C. Budd (2012), Resolution of sharp fronts in the presence of model error in variational data assimilation, *Quarterly Journal of the Royal Meteorological Society*.

Georgiou, T., J. Karlsson, and M. Takyar (2009), Metrics for power spectra: an axiomatic approach, *IEEE Transactions on Signal Processing*, *57*(3), 859–867.

Ghahramani, Z., and S. T. Roweis (1999), Learning nonlinear dynamical systems using an EM algorithm, *Advances in Neural Information Processing Systems*, pp. 431–437.

Gordon, N. J., D. J. Salmond, and A. F. Smith (1993), Novel approach to nonlinear/non-Gaussian Bayesian state estimation, *IEE Proceedings F (Radar and Signal Processing)*, *140*(2), 107–113.

Grant, M., S. Boyd, and Y. Ye (2012), CVX: Matlab software for disciplined convex programming, version 2.0 beta, *Recent Advances in Learning and Control*, pp. 95–110.

Griffith, A., and N. Nichols (2000), Adjoint Methods in Data Assimilation for Estimating Model Error, *Flow, Turbulence and Combustion*, *65*(3-4), 469–488.

Griffith, A. K., and N. Nichols ((1996)), Accounting for Model Error in Data Assimilation using Adjoint Methods, *In: Berz, M., Bischof, C., Corliss, G. and Griewank, A. (eds.) Computational Differentiation: Techniques, Applications and Tools. SIAM: Society for*

Industrial and Applied Mathematics, pp. 195–204.

Haber, E., T. Rehman, and A. Tannenbaum (2010), An efficient numerical method for the solution of the L_2 optimal mass transfer problem, *SIAM Journal on Scientific Computing*, *32*(1), 197–211.

Haker, S., L. Zhu, A. Tannenbaum, and S. Angenent (2004), Optimal mass transport for registration and warping, *International Journal of Computer Vision*, *60*(3), 225–240.

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

Hendricks Franssen, H., and W. Kinzelbach (2008), Real-time groundwater flow modeling using the ensemble Kalman filter: Join estimation of states and parameters and the filter inbreeding problem, *Water Resources Research*, *44*(9), W09408, doi: dpi:10.1029/2007WR006505.

Houser, P. R., W. J. Shuttleworth, J. S. Famiglietti, H. V. Gupta, K. H. Syed, and D. C. Goodrich (1998), Integration of soil moisture remote sensing and hydrologic modeling using data assimilation, *Water Resources Research*, *34*(12), 3405–3420.

Houser, P. R., G. J. M. De Lannoy, and J. P. Walker (2012), Hydrologic Data Assimilation, in *Approaches to Managing Disaster – Assessing Hazards, Emergencies and Disaster Impacts*, edited by J. P. Tiefenbacher, InTech.

Houtekamer, P., H. Mitchell, G. Pellerin, M. Buehner, M. Charron, L. Spacek, and B. Hansen (2005), Atmospheric data assimilation with an ensemble Kalman filter: Results with real observations, *Monthly Weather Review*, *133*(3), 604–620.

Johnson, C., N. N.K., and H. B.J. (2005), Very large inverse problems in atmospheric and ocean modeling, *International Journal of Numerical Methods in Fluids*, *47*, 759–771.

Kalman, R. E. (1960), A new approach to linear filtering and prediction problems, *Journal of Basic Engineering*, *82*(1), 35–45.

Kalnay, E., H. Li, T. Miyoshi, S.-C. Yang, and J. Ballabrera-Poy (2007), 4D–Var or ensemble Kalman filter?, *Tellus A*, *59*(5), 758–773.

Kantorovich, L. (1942), On the transfer of masses, in *Dokl. Akad. Nauk. SSSR*, vol. 37, pp. 227–229.

Kleist, D. T., D. F. Parrish, J. C. Derber, R. Treadon, W.-S. Wu, and S. Lord (2009), Introduction of the GSI into the NCEP Global Data Assimilation System, *Weather Forecasting*, *24*(6), 1691–1705.

Knott, M., and C. Smith (1984), On the optimal mapping of distributions, *Journal of Optimization Theory and Applications*, *43*(1), 39–49.

Kondrashov, D., C. Sun, and M. Ghil (2008), Data assimilation for a coupled ocean–atmosphere model. Part II: Parameter estimation, *Monthly Weather Review*, *136*(12), 5062–5076.

Koyama, H., and M. Watanabe (2010), Reducing forecast errors due to model imperfections using ensemble Kalman filtering, *Monthly Weather Review*, *138*(8), 3316–3332.

Kumar, S., R. Reichle, C. Peters-Lidard, R. Koster, X. Zhan, W. Crow, J. Eylander, and P. Houser (2008), A land surface data assimilation framework using the land information system: Description and applications, *Advances in Water Resources*, *31*(11), 1419–1432.

LeVeque, R. (1992), *Numerical methods for conservation laws*, Birkhäuser.

Liu, Y., and H. Gupta (2007), Uncertainty in hydrologic modeling: Towards an integrated data assimilation framework, *Water Resources Research*, *43 W07401*, W07,401, doi: dpi:10.1029/2006WR005756.

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 opportunities, *Hydrology and Earth System Sciences*, *16*(10), 3863–3887.

Lorenc, A. (1981), A global three-dimensional multivariate statistical interpolation scheme, *Monthly Weather Review*, *109*, 701–721.

Lorenc, A. C. (1986), Analysis methods for numerical weather prediction, *Quart. J. Roy. Meteor. Soc.*, *112*(474), 1177–1194.

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. J. Roy. Meteor. Soc.*, *126*(570), 2991–3012.

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 Methods in Fluids*, *40*(3-4), 435–444.

McLaughlin, D. (2002), An integrated approach to hydrologic data assimilation: interpolation, smoothing, and filtering, *Advances in Water Resources*, *25*, 1275–1286.

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

Moradkhani, H., C. M. DeChant, and S. Sorooshian (2012), Evolution of ensemble data assimilation for uncertainty quantification using the particle filter-markov chain monte carlo method, *Water Resources Research*, *48*(12), doi:10.1029/2012WR012144.

Nichols, N. (2003), Treating Model Error in 3-D and 4-D Data Assimilation, in *Data Assimilation for the Earth System*, *NATO Science Series*, vol. 26, edited by R. Swinbank, V. Shutyaev, and W. Lahoz, pp. 127–135, Springer Netherlands.

Noye, B., and H. Tan (1989), Finite difference methods for solving the two-dimensional advection–diffusion equation, *International Journal for Numerical Methods in Fluids*, *9*(1), 75–98.

Parrish, D. F., and J. C. Derber (1992), The National Meteorological Center's Spectral Statistical-Interpolation Analysis System, *Monthly Weather Review*, *120*(8), 1747–1763.

Reichle, R., W. Crow, and C. Keppenne (2008), An adaptive ensemble kalman filter for soil moisture data assimilation, *Water Resources Research*, *44*(3).

Reichle, R. H. (2008), Data assimilation methods in the Earth sciences, *Advances in Water Resources*, *31*(11), 1411–1418.

Reichle, R. H., D. B. McLaughlin, and D. Entekhabi (2002), Hydrologic data assimilation with the ensemble Kalman filter, *Monthly Weather Review*, *130*(1), 103–114.

Rubin, H., and J. F. Atkinson (2001), *Environmental fluid mechanics*, CRC Press.

Rubner, Y., C. Tomasi, and L. J. Guibas (2000), The earth mover's distance as a metric for image retrieval, *International Journal of Computer Vision*, *40*(2), 99–121.

Sasaki, Y. (1970), Some basic formalisms in numerical variational analysis, *Monthly Weather Review*, *98*(12), 875–883.

Sini, F., G. Boni, F. Caparrini, and D. Entekhabi (2008), Estimation of large–scale evaporation fields based on assimilation of remotely sensed land temperature, *Water Resources Research*, *44*(6), W06,410.

Talagrand, O. (1981), A study of the dynamics of four–dimensional data assimilation, *Tellus*, *33*(1), 43–60.

Trémolet, Y. (2006), Accounting for an imperfect model in 4D-Var, *Quarterly Journal of the Royal Meteorological Society*, *132*(621), 2483–2504.

van Leeuwen, P. J. (2010), Nonlinear data assimilation in geosciences: an extremely efficient particle filter, *Quarterly Journal of the Royal Meteorological Society*, *136*(653), 1991–1999.

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.

Villani, C. (2003), *Topics in optimal transportation*, vol. 58, American Mathematical Soc.

Vrugt, J. A., C. G. Diks, H. V. Gupta, W. Bouten, and J. M. Verstraten (2005), Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation, *Water Resour. Res.*, *41*(1).

Weickert, J., and C. Schnörr (2001), A theoretical framework for convex regularizers in PDE-based computation of image motion, *International Journal of Computer Vision*, *45*(3), 245–264.

Zupanski, D. (1997a), A General Weak Constraint Applicable to Operational 4DVAR Data Assimilation Systems, *Monthly Weather Review*, *125*(9), 2274–2292.

Zupanski, D. (1997b), A general weak constraint applicable to operational 4DVAR data assimilation systems, *Monthly Weather Review*, *125*(9), 2274–2292.

Zupanski, D., and M. Zupanski (2006), Model error estimation employing an ensemble data assimilation approach, *Mon. Wea. Rev.*, *134*(5), 1337–1354.

Figure 1. Left panel: initial state $\mathbf{x}_0$, true $\mathbf{x}_1$, and incorrect states $\tilde{\mathbf{x}}_1^{(i)}$, $i \in \{1,2,3,4\}$ obtained by an erroneous model, here, a biased estimate of the shift coefficient at time $t = 1$; Right panel: quadratic norm $\left\|\mathbf{x}_1 - \tilde{\mathbf{x}}_1^{(i)}\right\|_2^2$ and OMT-metric $\mathcal{T}\left(\mathbf{x}_1, \tilde{\mathbf{x}}_1^{(i)}\right)$ 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).

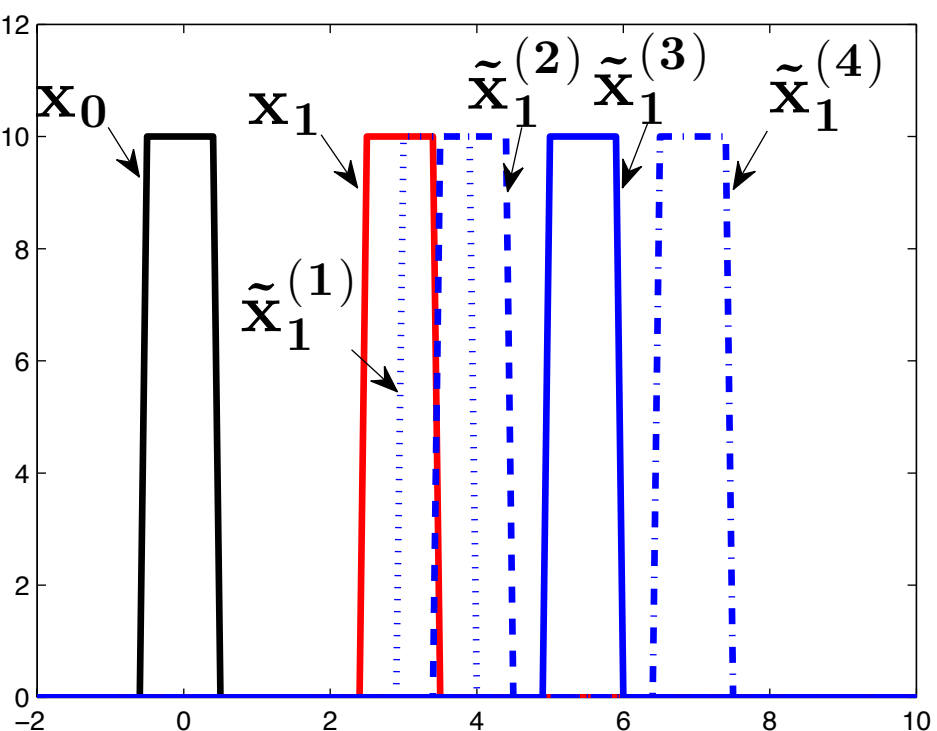
Figure 2. Left panel: true density $\mathbf{x}$ (in the solid line) and two oppositely biased and noisy densities $\mathbf{x}_b$ (in the dashed line) and $\mathbf{y}$ (in the dash-dotted line). Right panel: true density $\mathbf{x}$ (in the solid blue line) and the ℓ_2 ($\hat{\mathbf{x}}^{\ell_2}$ in the dashed green line) and OMT-estimates ($\hat{\mathbf{x}}^{\text{omt}}$ in dash-dotted green line), respectively.

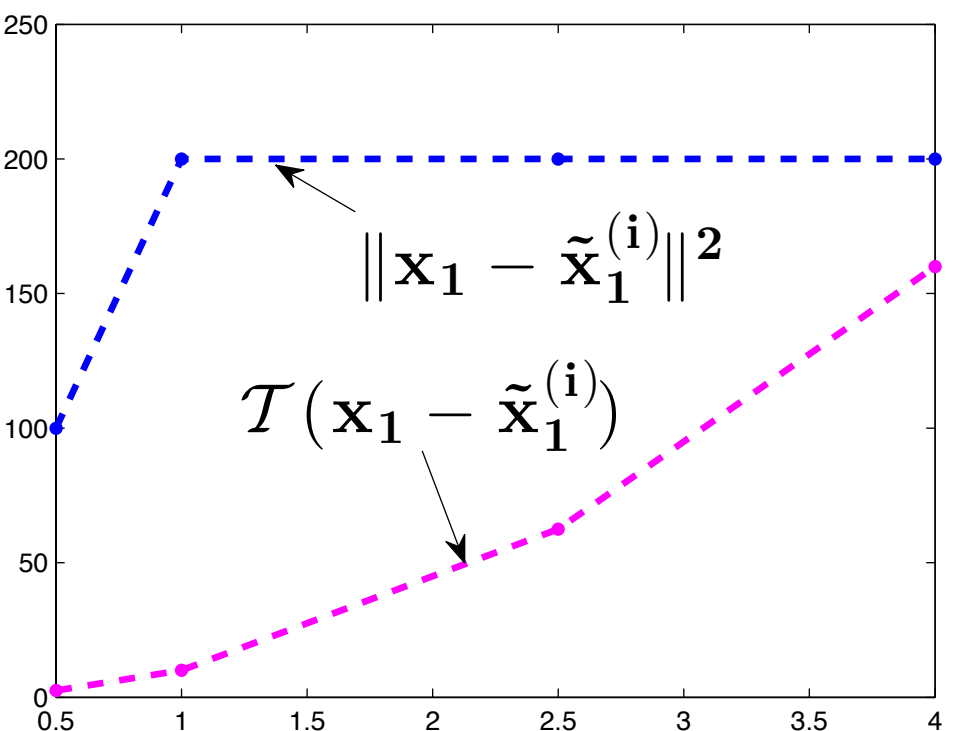
Figure 3. Left panel: initial state $\mathbf{x}_0$ (solid line), background state $\mathbf{x}_b$ (dotted line), and two model simulated states $\mathbf{x}_1$, $\mathbf{x}_2$ at time steps $t = 1, 2$ (dash-dotted and dashed lines, respectively). The simulated states $\mathbf{x}_1, \mathbf{x}_2$ were produced from $\mathbf{x}_0$ using the advection diffusion model with parameters $u = 0.75$ and $D = 0.02$, while $\mathbf{x}_b$ was taken to be a lagged and diffused variant of $\mathbf{x}_0$. Right panel: noisy observations $\mathbf{y}_0$, $\mathbf{y}_1$, and $\mathbf{y}_2$ at time steps $t = 0, 1, 2$ produced from $\mathbf{x}_0$, $\mathbf{x}_1$, $\mathbf{x}_2$, respectively. The components of the process and observation noises were assumed Gaussian with zero mean and variance $R = 0.2$.

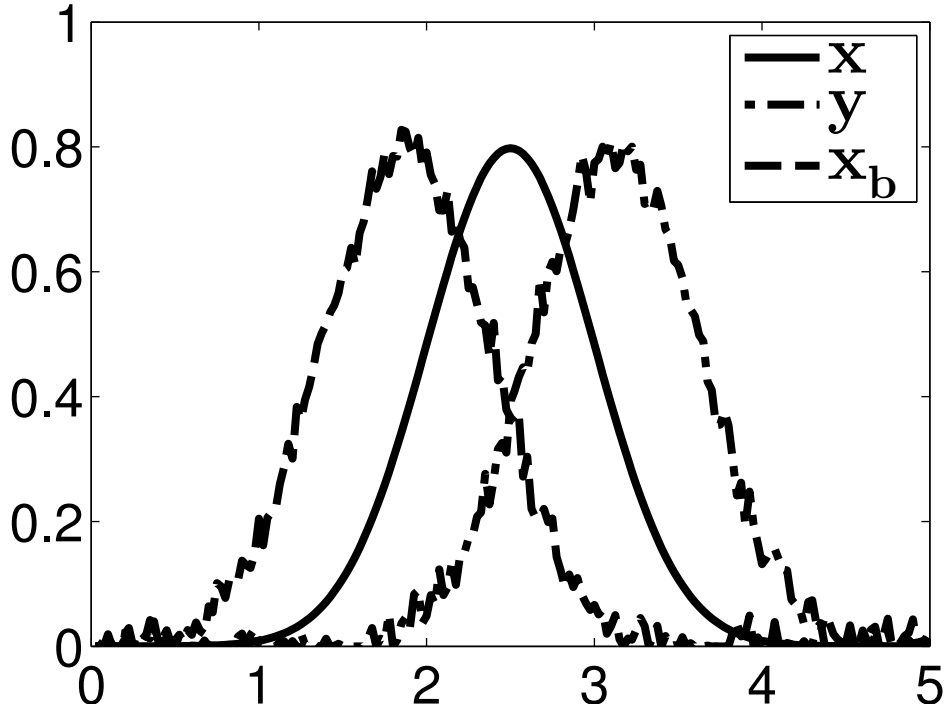
Figure 4. From left to right: true states $\mathbf{x}_t$ (solid blue line) and estimated states $\hat{\mathbf{x}}_t$, at time steps $t = 0, 1, 2$ in successive panels, using ℓ_2 and OMT and the erroneous advection diffusion model with parameters $\tilde{u} = 0.5$ and $\tilde{D} = 0.02$. At $t = 0$, ℓ_1 regularization has also been used to promote the sparseness of the state in the derivative space. The accuracy of the OMT estimates is notable.

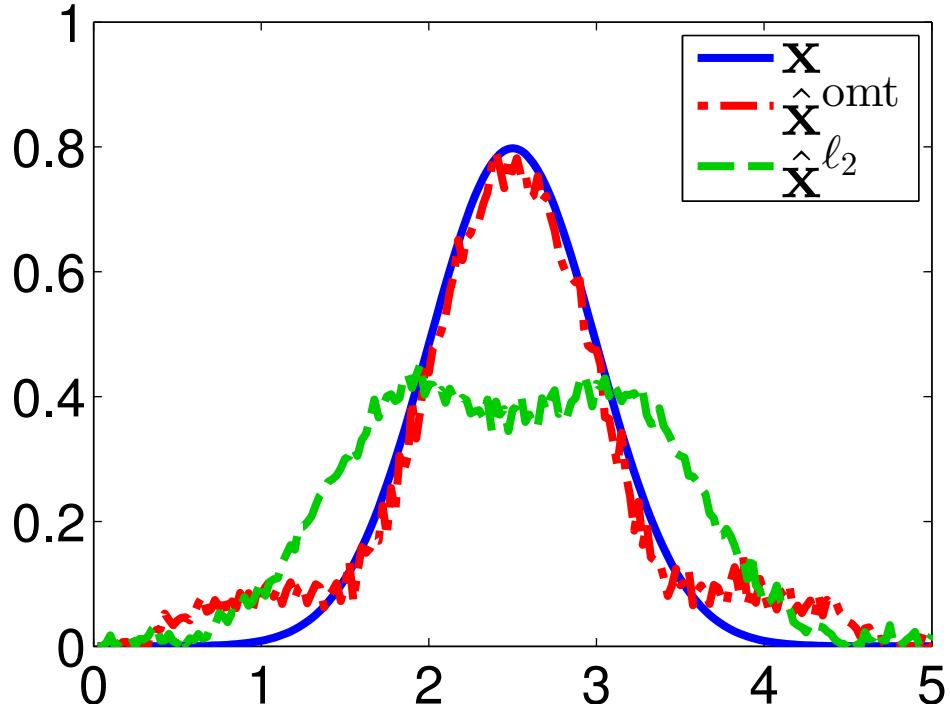
Figure 5. From left to right: the biased background state X_b (produced as a prediction from a previous time step using the biased model $\tilde{u}_x = \tilde{u}_y = 0.1$ and $\tilde{D}_x = \tilde{D}_y = 0.015$) and the observations Y_0, Y_1 and Y_2 at time steps $t = 0, 1, 2$. The observations $Y_k = HX_kH^T + V_k$ 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.

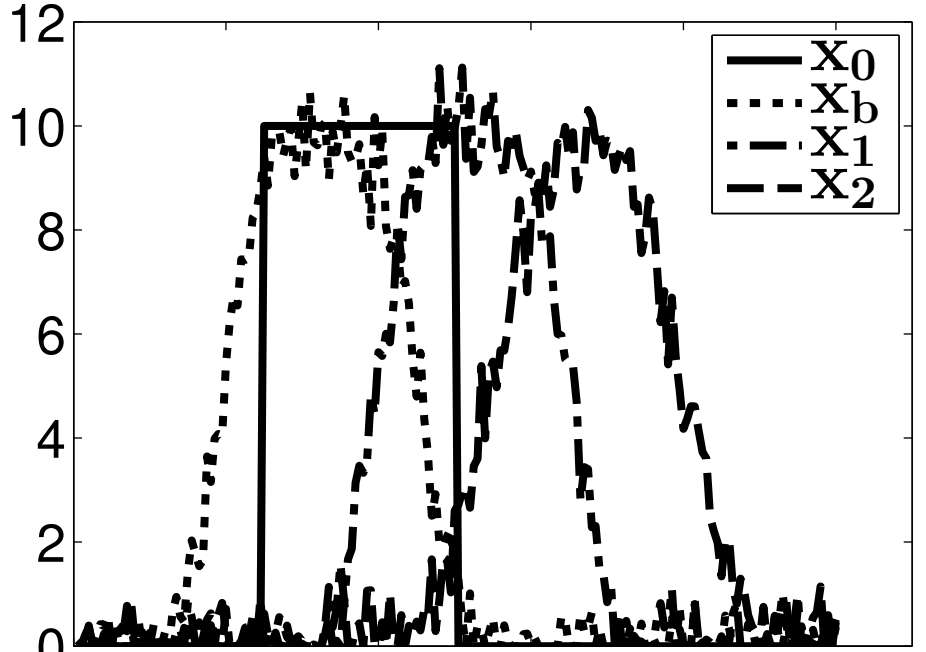
Figure 6. First row panels, from left to right: true state X_0 and its estimates via weak–constraint 4D-Var plus ℓ_1-norm regularization, and via the proposed OMT–based methodology, based on the biased background state X_b and the observations Y_k (shown in Figure 5). Note that in all cases the biased 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.

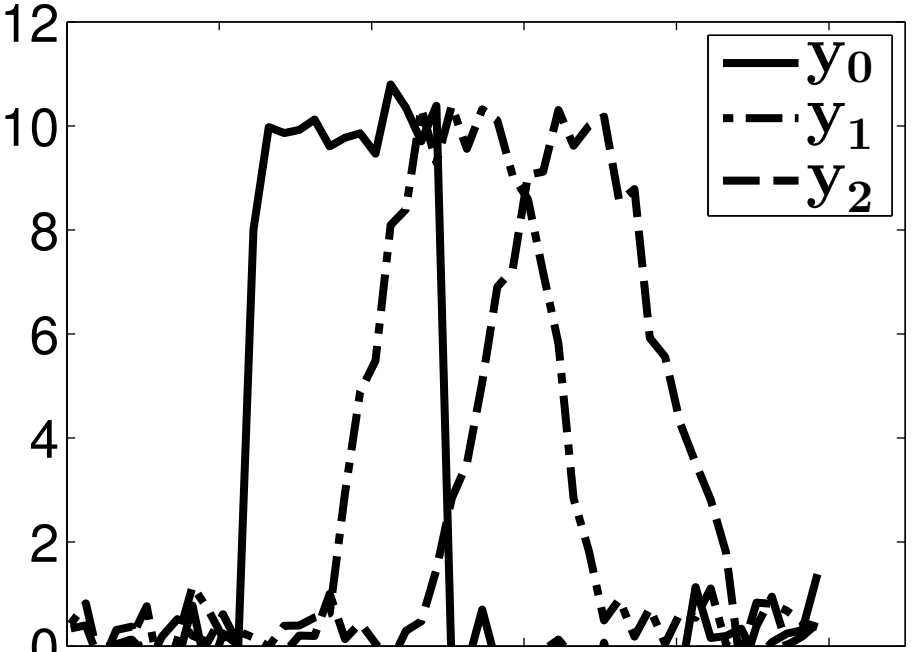

$\mathbf{x_0}$
$\mathbf{x_1}$
$\tilde{\mathbf{x}}_1^{(1)}$
$\tilde{\mathbf{x}}_1^{(2)}$
$\tilde{\mathbf{x}}_1^{(3)}$
$\tilde{\mathbf{x}}_1^{(4)}$
12
10
8
6
4
2
0
−2
0
2
4
6
8
10

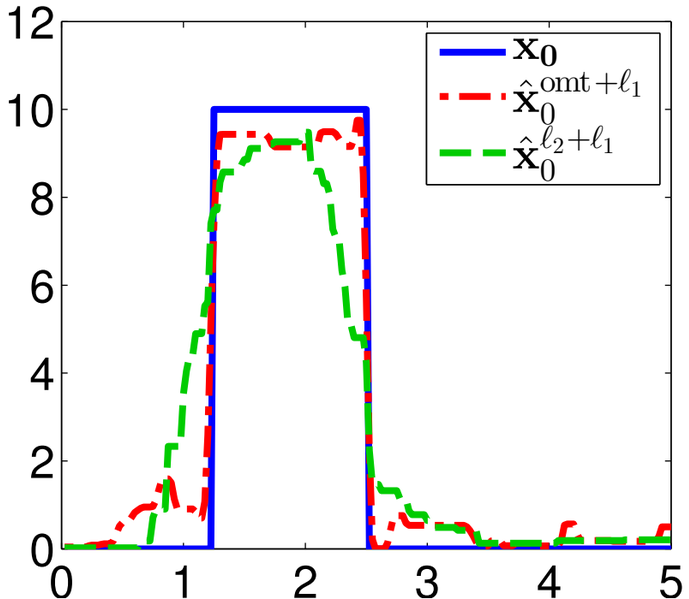

$\|\mathbf{x_1} - \tilde{\mathbf{x}}_1^{(i)}\|^2$
$\mathcal{T}\left(\mathbf{x_1} - \tilde{\mathbf{x}}_1^{(i)}\right)$
250
200
150
100
50
0
0.5
1
1.5
2
2.5
3
3.5
4

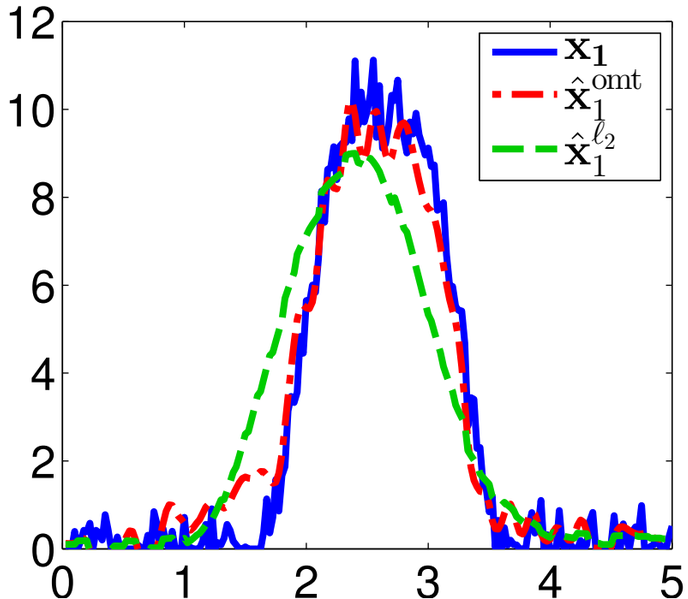

x
y
x_b
0
0.2
0.4
0.6
0.8
1
0
1
2
3
4
5

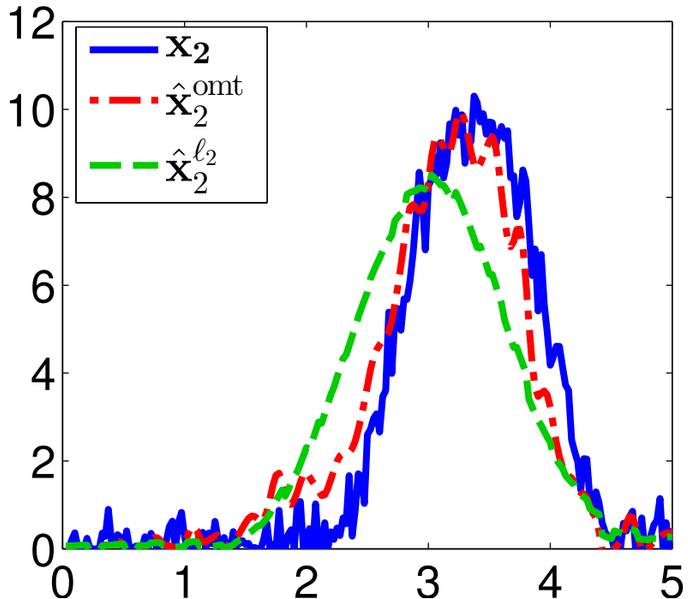

x
$\hat{x}^{omt}$
$\hat{x}^{\ell_2}$
0
0.2
0.4
0.6
0.8
1
0
1
2
3
4
5

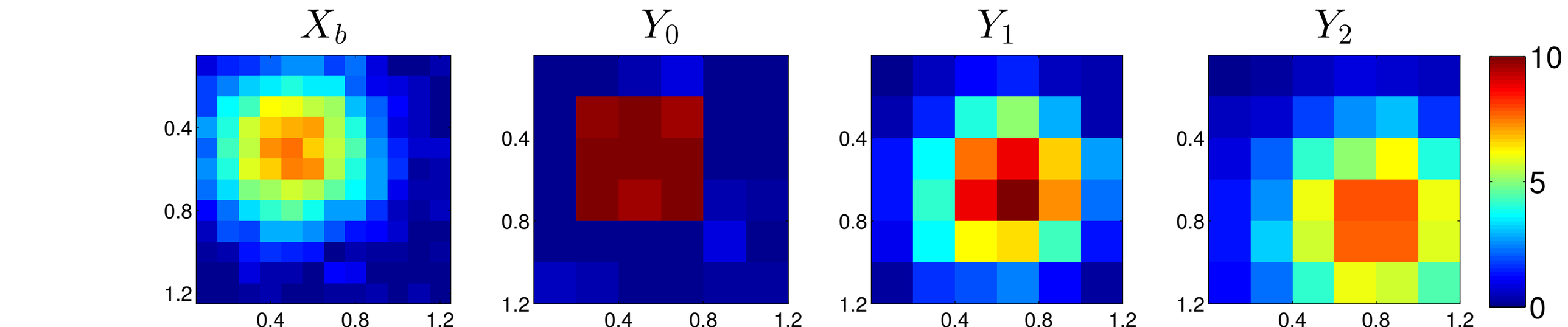

12
10
8
6
4
2
0
0
1
2
3
4
5
x_0
x_b
x_1
x_2

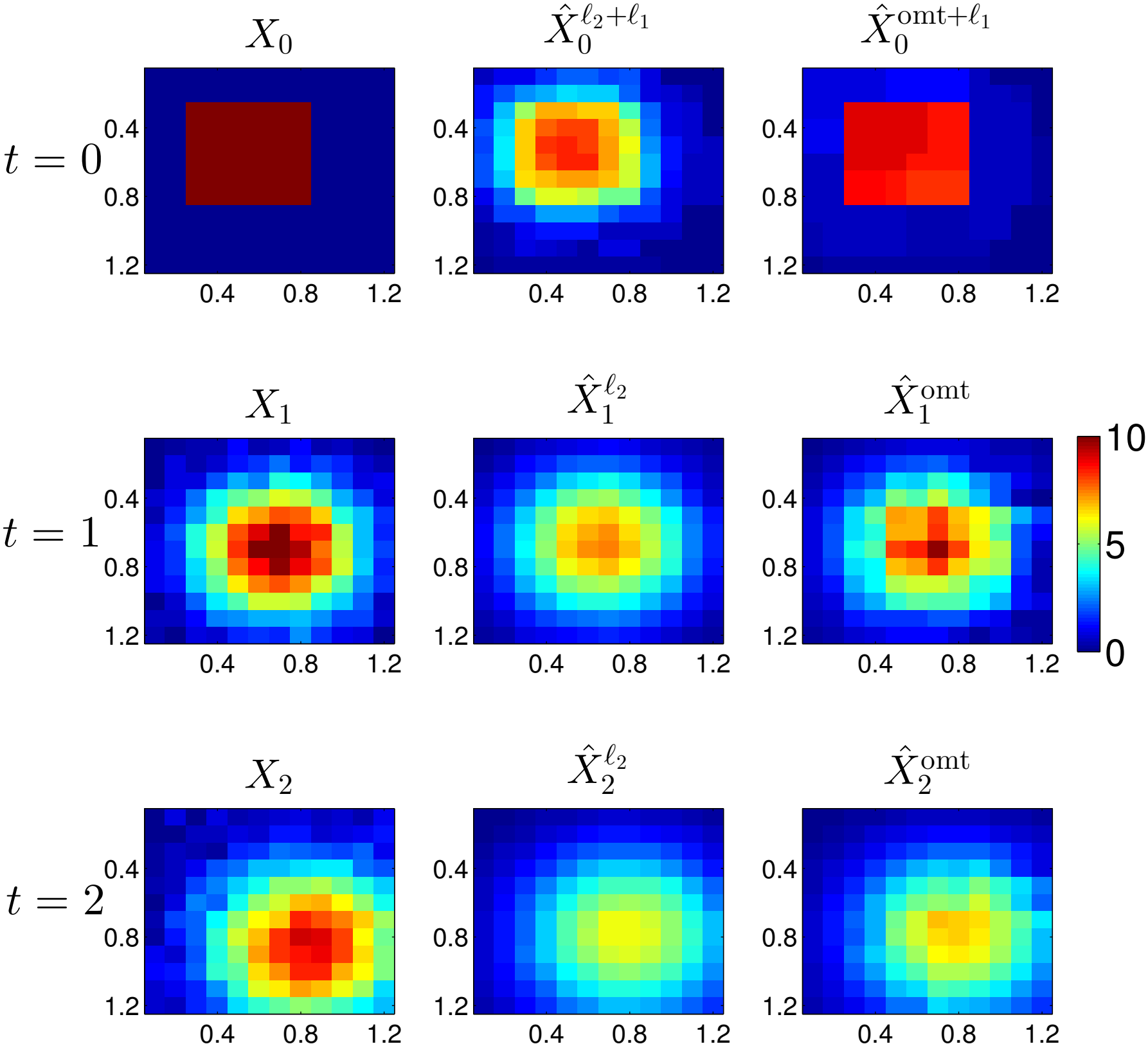

12
10
8
6
4
2
0
0
1
2
3
4
5
y_0
y_1
y_2


$\mathbf{x_0}$
$\hat{\mathbf{x}}_0^{\text{omt}+\ell_1}$
$\hat{\mathbf{x}}_0^{\ell_2+\ell_1}$
12
10
8
6
4
2
0
0
1
2
3
4
5


$\mathbf{x_1}$
$\hat{\mathbf{x}}_1^{\text{omt}}$
$\hat{\mathbf{x}}_1^{\ell_2}$
12
10
8
6
4
2
0
0
1
2
3
4
5


$\mathbf{x_2}$
$\hat{\mathbf{x}}_2^{\text{omt}}$
$\hat{\mathbf{x}}_2^{\ell_2}$
12
10
8
6
4
2
0
0
1
2
3
4
5


X_b
Y_0
Y_1
Y_2
0.4
0.8
1.2
10
5
0


X_0
$\hat{X}_0^{\ell_2+\ell_1}$
$\hat{X}_0^{\mathrm{omt}+\ell_1}$
$t=0$
X_1
$\hat{X}_1^{\ell_2}$
$\hat{X}_1^{\mathrm{omt}}$
$t=1$
X_2
$\hat{X}_2^{\ell_2}$
$\hat{X}_2^{\mathrm{omt}}$
$t=2$
0.4
0.8
1.2
10
5
0