Regularized Variational Data Assimilation for Bias Treatment using the Wasserstein Metric

Sagar K. Tamang*, Ardeshir M. Ebtehaj

Saint Anthony Falls Laboratory and Department of Civil, Environmental and Geo-Engineering, University of Minnesota-Twin Cities, Minneapolis, Minnesota

Dongmian Zou, Gilad Lerman

School of Mathematics, University of Minnesota-Twin Cities, Minneapolis, Minnesota

*Corresponding author address: Department of Civil, Environmental and Geo-Engineering, University of Minnesota-Twin Cities, Minneapolis, Minnesota

E-mail: taman011@umn.edu
ABSTRACT

This paper presents a new variational data assimilation (VDA) approach for the formal treatment of bias in both model outputs and observations. This approach relies on the Wasserstein metric stemming from the theory of optimal mass transport to penalize the distance between the probability histograms of the analysis state and an a priori reference dataset, which is likely to be more uncertain but less biased than both model and observations. Unlike previous bias-aware VDA approaches, the new Wasserstein metric VDA (WM-VDA) dynamically treats systematic biases of unknown magnitude and sign in both model and observations through assimilation of the reference data in the probability domain and can fully recover the probability histogram of the analysis state. The performance of WM-VDA is compared with the classic three-dimensional VDA (3D-Var) scheme on first-order linear dynamics and the chaotic Lorenz attractor. Under positive systematic biases in both model and observations, we consistently demonstrate a significant reduction in the forecast bias and unbiased root mean squared error.
1. Introduction

The predictive accuracy of the Earth System Model (ESM) relies on a series of differential equations that are often sensitive to their initial conditions. Even a small error in estimates of their initial conditions can lead to large forecast uncertainties. In short-to-medium range forecast systems, an open-loop run of coupled land and weather models often diverges from the true states and show low forecast skills as the error keeps accumulating over time (Charney, 1951; Kalnay et al., 2007). To extend the forecast lead time, the science of data assimilation (DA) attempts to use the information content of the observations for improved estimates of the ESMs initial conditions thus reducing their forecast uncertainties (Leith, 1993; Kalnay, 2003). The DA methods often involve iterative cycles at which the observations are optimally integrated with the previous time forecasts (background states) to obtain an a posteriori estimate of the initial conditions (analysis state) with reduced uncertainty in a Bayesian setting (Rabier, 2005; Asch et al., 2016). In the literature, two major categories of DA methodologies exist, namely filtering and variational methods (Law and Stuart, 2012). Advanced approaches such as hybrid DA schemes are also being developed, which aim to combine and take advantage of unique benefits of the two DA classes (Wang et al., 2008; Lorenc et al., 2015).

Although, both classic DA approaches have been widely used in land and weather forecast systems, they are often based on a strict underlying assumption that the error is drawn from a zero-mean Gaussian distribution, which is not always realistic. Bias exists in land-atmosphere models mainly due to under-representation of the governing laws of physics and erroneous boundary conditions. Observation bias also exists largely due to systematic errors in sensing systems and retrieval algorithms represented by the observation operator in DA systems (Dee, 2003, 2005). From a mathematical point of view, bias is simply the expected value of the error that can be removed if
the ground truth of the process was known, which is not often feasible in reality. The problem is often exacerbated due to difficulty in attribution of the bias to either model or observations and/or both. At the same time, point-scale observations such as those from in-situ gauges and radiosondes are often considered to be more close to the ground truth, however, their assimilation into gridded model outputs is not straightforward due to the existing scale gaps.

Bias correction strategies in DA systems mainly fall under two general categories: (i) dynamic bias-aware and (ii) re-scaling bias correction schemes. Apart from these two general categories, machine learning techniques have also been developed to learn relationships between observations and ancillary variables for bias correction (Jin et al., 2019). The dynamic bias-aware schemes make prior assumptions about the nature of the bias and attribute it either to model or observations which may not be realistic as both models and observations suffer from systematic errors. Early attempts to dynamically treat model biases are based on a two-step bias estimation and correction approach, which is applied prior to the analysis step (Dee and Da Silva, 1998; Radakovich et al., 2001; Chepurin et al., 2005). Variants of bias-aware Kalman filter for colored noise (Drécourt et al., 2006) and a weak constrained four-dimensional VDA (4D-Var) (Zupanski, 1997) have also been proposed to account for non-zero mean model errors. At the same time, there exists another class of dynamic observational bias correction techniques that rely on variants of variational bias-correction (VarBC) method, which makes an a priori estimate of bias and dynamically update it using innovation information (Auligné et al., 2007; Dee and Uppala, 2009; Zhu et al., 2014). Apart from VarBC, more recently, a new approach is proposed to treat observation biases by iterative updates of the observation operator (Hamilton et al., 2019). A body of research also has been devoted for simultaneously treating model and observation biases using multi-stage hybrid filtering technique (Pauwels et al., 2013). However, the above schemes still lack the ability to leverage climatologically unbiased information from reference observations (e.g., in situ data) and
has not yet been tested for effective bias correction in chaotic systems. More importantly, the developed schemes largely focus to retrieve an unbiased expected value of the forecast and remain limited to characterization of the second-order forecast uncertainty.

The re-scaling techniques do not make any explicit assumptions about relative accuracy of the model and observation system (Reichle and Koster, 2004; Crow et al., 2005; Reichle et al., 2007; Kumar et al., 2009; Reichle et al., 2010; Liu et al., 2018). This family of methods often involves mapping the observations onto the model space by matching their Cumulative Distribution Function (CDF). While the CDF-matching technique is comparatively easier in implementation than the dynamic approach and prevents any numerical instabilities in model simulations, it implicitly assumes that model forecasts are unbiased and partly ignores the information content of observations. For example, if our observations were less biased than the model outputs, this approach basically fails to effectively remove the bias. Furthermore, it is a static scheme and there exists no formal way to extend CDF-matching scheme to dynamically account for changes in the bias (Kumar et al., 2012) and its seasonality (De Lannoy et al., 2016).

Conceptually, CDF-matching techniques move probability masses from one distribution to another. To transform the static CDF-matching to a dynamic scheme, there are two key questions that we aim to answer: Can we quantify the movement of probability masses as a cost through a convex metric? How can this cost be employed to assimilate relatively unbiased in situ data for dynamic bias correction in the VDA framework?

The Wasserstein metric (WM) (Villani, 2008; Santambrogio, 2015), also known as the Earth Mover’s distance (Rubner et al., 2000) stems from the theory of optimal mass transport (OMT) (Monge, 1781; Kantorovich, 1942; Villani, 2003), which provides a concrete ground to compare probability measures (Brenier, 1991; Gangbo and McCann, 1996; Benamou et al., 2015; Chen et al., 2017, 2018a,b). Specifically, this distance metric can quantify the dissimilarity between
two probability histograms in terms of the amount of “work” done during displacement of probability masses between them. Thus, we hypothesize that inclusion of such metric in the VDA cost function can reduce analysis biases. The rationale is that the work done during displacement of the probability masses, is not only a function of the shape of probability histograms but also the difference between their central positions, as described in Section c. The use of the Wasserstein metric in DA has been explored previously (Ning et al., 2014; Feyeux et al., 2018). Ning et al. (2014) introduced the concept of OMT in classical VDA framework and demonstrated that the bias in the background state results in an unrealistic bimodal distribution of the analysis state. However, the study was conducted on linear systems only for model bias correction without accounting for any form of observation biases. Feyeux et al. (2018) proposed to fully replace the quadratic costs in the classic VDA by the Wasserstein metric. Even though, the latter approach extends the classic VDA beyond a minimum mean squared error approximation, it does not provide any road map for bias correction, which is the central focus of this paper.

This paper presents a new VDA approach through regularizing the classic VDA problem with cost associated with the Wasserstein metric, hereafter referred to as the Wasserstein Metric VDA (WM-VDA). Unlike previous VDA techniques, WM-VDA treats unknown biases of different magnitudes and signs both in the model dynamics and observations. To that end, WM-VDA needs to be informed by an a priori reference distribution or histogram (e.g., from in situ data) that encodes the space-time variability of the state variables of interest in probability domain. This a priori histogram must be less biased but could exhibit larger higher-order uncertainties than the observations and model forecasts. More importantly, unlike classic DA methods, the WM-VDA allows full recovery of the probability histogram of the analysis state in the probability domain, which can lead to forecast uncertainty quantification beyond second-order statistics. The idea is tested on a first-order linear dynamical system as a test-bed and the chaotic Lorenz-63 (Lorenz, 1963).
attractor that represents nonlinear dynamics of convective circulation in a shallow fluid layer. The results demonstrate that the presented approach is capable in preserving the geometric shape of the distribution of analysis state when both the background state and observations are systematically biased and extend the forecast skills by controlling the propagation of bias in the phase space of a highly chaotic system.

The paper is organized as follows: Section 2 discusses the concept of classic VDA focusing on the 3D-Var. In this section, a summary on the theory of OMT and the Wasserstein metric is also provided. The mathematical formulation of the proposed WM-VDA is explained in Section 3. Section 4 implements the WM-VDA on a first-order linear and the nonlinear Lorenz-63 dynamic systems. The results are interpreted and compared with the 3D-Var and CDF-matching technique. Summary and concluding remarks are presented in Section 5.

2. Methodology

a. Notations

Throughout, small and capital boldface letters are reserved for representation of $m$-element column vectors $\mathbf{x} \in \mathbb{R}^m$ and $m$-by-$n$ matrices $\mathbf{X} \in \mathbb{R}^{m \times n}$, $\mathbf{1}_m$ is an $m$-element vector of ones and $\mathbf{I}_m$ denotes an $m \times m$ identity matrix. A 1-D state variable of interest $\mathbf{x} \in \mathbb{R}$ is represented by a probability vector $p_x = (p_{x_1}, \ldots, p_{x_k})^T$ supported on $k$ points $x_1, \ldots, x_k$, such that $\mathbf{x} = \sum_k p_{x_k} \delta_{x_k}$, where $\delta_{x_k}$ is the Dirac function at $x_k$ and $(\cdot)^T$ denotes the transposition operator. For the state $\mathbf{x} \in \mathbb{R}^m$, this linear expectation operator is represented as $\mathbf{x} = \mathbf{X} p_x$, where the support point and their associated probability of occurrences are properly concatenated in $\mathbf{X} \in \mathbb{R}^{m \times k}$ and $p_x \in \mathbb{R}^k$, respectively. $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$ denotes that $\mathbf{x}$ is drawn from a Gaussian distribution with mean $\mu$ and covariance
\[ \Sigma, \text{ and square of the weighted } \ell_2\text{-norm of } x \text{ is represented as } \|x\|_{B^{-1}}^2 = x^T B^{-1} x, \text{ where } B \text{ is a positive definite matrix.} \]

\textit{b. Classic 3D-Var}

In three-dimensional VDA (3D-Var \textit{Lorenc} [1986]), the analysis state is a weighted average of the background state and observations with the weights defined by their respective error covariance matrices. Specifically, let us assume that the \(m\)-element state variable at time step \(t = i + 1\) is denoted by \(x_{i+1} \in \mathbb{R}^m\) with the following stochastic dynamics:

\[ x_{i+1} = \mathcal{M}(x_i) + \omega_i, \quad (1) \]

where \(\mathcal{M} : \mathbb{R}^m \to \mathbb{R}^m\) is the nonlinear model operator evolving the state from \(x_i\) to \(x_{i+1}\) and \(\omega_i \sim \mathcal{N}(0, B) \in \mathbb{R}^m\) is the model error and the expected background state at \((i+1)\)th time step is \(x_b = \mathcal{M}(x_i)\).

Additionally, the \(n\)-element observation vector available at discrete time step \(i\) is denoted by \(y_i \in \mathbb{R}^n\) and is related to the true state as follows:

\[ y_i = \mathcal{H}(x_i) + v_i, \quad (2) \]

where \(\mathcal{H} : \mathbb{R}^m \to \mathbb{R}^n\) is a nonlinear operator mapping the state space to the observation space and \(v_i \sim \mathcal{N}(0, R) \in \mathbb{R}^n\) is the observation error.

In a 3D-Var setting, the cost function is comprised of two weighted Euclidean distance of the unknown true state from the background state \(x_b\) and the observation \(y\) such that,

\[ J_{3D}(x_0) = \|x_0 - x_b\|_{B^{-1}}^2 + \|y - \mathcal{H}(x_0)\|_{R^{-1}}^2. \quad (3) \]

Assuming that the nonlinear observation operator can be well approximated linearly such that \(\mathcal{H}(x_i) = \mathbf{H} x_i\), the estimation of the analysis state \(x_a \in \mathbb{R}^m\) at any time step amounts to minimizing
the quadratic cost function as follows:

\[
x_a = \arg \min_{x_0} J_{3D}(x_0)
= \arg \min_{x_0} \left\{ \|x_0 - x_b\|^2_{B^{-1}} + \|y - Hx_0\|^2_{R^{-1}} \right\}.
\]

Since the minimization problem presented in Equation 4 is convex, the local minimum is the global minimum. Thus, by setting the first-order derivative to zero, the analysis state \(x_a\) is obtained as

\[
x_a = (H^T R^{-1} H + B^{-1})^{-1} (H^T R^{-1} y + B^{-1} x_b),
\]

where the analysis error covariance \(P_a\) is the inverse of the Hessian of Equation 3 (Daley, 1993):

\[
P_a = (H^T R^{-1} H + B^{-1})^{-1}.
\]

Clearly the above representations are only valid for a linear observation operator and shall be considered as an approximate for the nonlinear case. It is also worth mentioning that under the assumption of zero-mean Gaussian errors and linear observation operator, the obtained analysis state, as derived in Equation 5, can be interpreted as the minimum variance unbiased estimator or the maximum \textit{a posteriori} estimator. It is also worth noting that in this setting the results of the 3D-Var is equivalent to the update equations used in standard Kalman filter; however, become suboptimal for non-Gaussian errors and/or a nonlinear observation operator (Courtier et al., 1994).

c. Optimal Mass Transport

The application of the theory of optimal mass transport (OMT) pioneered by Gaspard Monge (Monge, 1781) seems to be a natural extension of the CDF-matching techniques for dynamic bias correction in VDA. The theory was first conceptualized to minimize the total amount of work in transportation of materials between two locations. Recent advances of the theory has provided a fertile ground for comparison of the probability measures (Brenier, 1991; Villani, 2003) and has
been extensively studied and widely applied in the field of signal processing (Kolouri et al., 2017; Motamed and Appelo, 2018), image retrieval (Rubner et al., 2000; Li et al., 2013), and in analyzing misfit in seismic signals (Engquist and Froese, 2013).

Let \( p_x = (p_{x_1}, \ldots, p_{x_k})^T \in \mathbb{R}^k \) and \( p_z = (p_{z_1}, \ldots, p_{z_l})^T \in \mathbb{R}^l \) represent the probability vectors associated with a source and a target histogram supported on vectors \( x_1, \ldots, x_k \) and \( z_1, \ldots, z_l \), respectively. In Monge formulation, the problem involves seeking a surjective optimal transport map \( T: \{x_1, \ldots, x_k\} \rightarrow \{z_1, \ldots, z_l\} \) that moves probability mass from each discrete point \( x_i \) on source probability histogram to a “single” point \( z_j \) on target probability histogram, where \( i = 1, \ldots, k \) and \( j = 1, \ldots, l \), such that the total cost of transportation is minimized:

\[
\begin{align*}
\text{minimize} \quad & \sum_i c(x_i, T(x_i)) \\
\text{subject to} \quad & p_{z_j} = \sum_{i: T(x_i) = z_j} p_{x_i},
\end{align*}
\]

where \( c(x_i, T(x_i)) \) is the transportation cost between points \( x_i \) and \( T(x_i) \) and the constraint warrants the mass conservation principle. Essentially, Monge formulation of OMT problem is non-convex and becomes a combinatorial NP-hard problem, for which the transport map \( T(\cdot) \) does not exist – when target probability histogram has more support points than the source histogram (i.e. \( l > k \)) (Peyré et al., 2019).

Kantorovich (1942) proposed a convex relaxation of the Monge OMT problem through probabilistic consideration of transport in which mass at any source point \( x_i \) can be split and transported across several target points \( z_j \). A schematic of the difference between Monge and Kantorovich formulations of the OMT is shown in Figure 1. Let us assume that \( U \in \mathbb{R}^{k \times l} \) represents the so-called transportation plan matrix, where the element \( u_{ij} \) describes the probability mass being transferred from point \( x_i \) to point \( z_j \). Here, \( C \in \mathbb{R}^{k \times l} \) represents the transportation or the ground cost matrix, where the element \( c_{ij} = |x_i - z_j|^p \) is the cost of transporting probability masses from \( x_i \) to \( z_j \) and
\( p \) is a positive exponent. Assuming such a ground cost, turns the Kantorovich formulation to the \( p \)-Wasserstein (\( \mathcal{W}_p \)) distance or metric, which seeks to minimize the total amount of work done in transporting probability masses from \( \mathbf{p}_x \) to \( \mathbf{p}_z \) as follows:

\[
\mathcal{W}_p(\mathbf{p}_x, \mathbf{p}_z) := \left( \min \sum_{i,j} c_{ij} u_{ij} \right)^{1/p} = \left( \min \text{tr}(\mathbf{C}^T \mathbf{U}) \right)^{1/p}
\]

subject to \( u_{ij} \geq 0 \)

\[
\mathbf{U}^T \mathbf{e}_l = \mathbf{p}_x
\]

\[
\mathbf{U}^T \mathbf{e}_k = \mathbf{p}_z.
\]

The first constraint assures that the transported probability masses are non-negative and the second and third constraints warrant that the transportation follows conservation of mass. Thus, the Wasserstein metric between two probability histograms is the minimum cost to match them through a transportation plan matrix. Unlike the Euclidean distance that penalizes the second order statistics of error, the Wasserstein metric penalizes the misfit between the shape of the histograms and increases monotonically with a shift between central position of the histograms (Ning et al., 2014), enabling it to naturally penalize the bias. More specifically, for \( p = 2 \), it can be shown that 

\[
\mathcal{W}_2^2(\mathbf{p}_x, \mathbf{p}_z) = \mathcal{W}_2^2(\tilde{\mathbf{p}}_x, \tilde{\mathbf{p}}_z) + \| \mu_x - \mu_z \|_2^2,
\]

where \( \tilde{\mathbf{p}}_x \) and \( \tilde{\mathbf{p}}_z \) are the centred zero-mean probability masses and \( \mu_x \) and \( \mu_z \) are the mean values.

3. Regularization of VDA through the Wasserstein Metric

Regularization techniques have been used to reduce uncertainty of the analysis state (Wahba and Wendelberger, 1980; Lorenc, 1986) with isolated singularities (Ebtehaj et al., 2014) —focusing on precipitation convective cells (Ebtehaj and Foufoula-Georgiou, 2013), sharp transitions in weather fronts (Freitag et al., 2010) and sea ice thickness (Asadi et al., 2019). This paper proposes to add the cost of the square of the 2-Wasserstein distance as a regularization term to the classic VDA.
scheme, referred to as WM-VDA. The consideration of 2-Wasserstein distance also ensures keeping the DA problem convex. This approach not only penalizes the background and observation error in a least-squares sense but also takes into consideration the mismatch between the probability distributions of the analysis state and a “relatively unbiased” a priori reference probability histogram. For example, this histogram can be obtained from the soil moisture gauge measurements, sea ice buoy data or atmospheric radiosondes. Such in situ measurements are often relatively less biased than both remotely sensed satellite retrievals and the background state, but can exhibit larger higher-order uncertainties, over a sufficiently large window of time or space (Villarini et al., 2008).

Specifically, let us assume that \( p_x = (p_{x1}, \ldots, p_{xkm})^T \in \mathbb{R}^{km} \) and \( p_{xr} = (p_{xr1}, \ldots, p_{xrkm})^T \in \mathbb{R}^{km} \) represent the vertically concatenated probability vectors of the analysis state and the a priori reference data supported on the “known” matrix \( X = [x_1 | \ldots | x_{km}] \in \mathbb{R}^{m \times km} \) where, \( k \) represents the number of domain discretization of the state variable in each dimension and \( x_i \in \mathbb{R}^m \) represents the \( m \)-dimensional vertically concatenated vector of the support points for their associated probability vectors. The WM-VDA cost function is then defined as

\[
J_{WM-VDA}(x, p_x) = \|x - x_b\|_{B^{-1}}^2 + \|y - Hx\|_{R^{-1}}^2 + \lambda J_{W2}(p_x, p_{xr}),
\]

where \( J_{W2}(p_x, p_{xr}) \) represents the transportation cost associated with the square of the 2-Wasserstein distance between the two probability histograms and \( \lambda \) is a non-negative regularization parameter, which balances a trade-off between the Euclidean and the Wasserstein cost. There is no closed form solution to determine this hyper-parameter; therefore, it should be estimated empirically through cross-validation experiments.
The analysis state $x$ at initial time is an expected value that can be represented as $x = Xp_x$. Thus, Equation 9 can be expanded as follows:

$$J_{WM-VDA}(p_x) = \|Xp_x - x_b\|_B^{-1} + \|y - HXp_x\|_R^{-1} + \lambda J_w(p_x, p_x_r).$$

(10)

From the second mass constraint in Equation 8, we have $p_x = U\kappa^{km}$ and setting $J_w(p_x, p_x_r) = \text{tr}(C^T U)$, the above cost function can be expressed in terms of the transportation cost matrix $U$:

$$J_{WM-VDA}(U) = \|XU\kappa^{km} - x_b\|_B^{-1} + \|y - HXU\kappa^{km}\|_R^{-1} + \lambda \text{tr}(C^T U).$$

(11)

Let us assume that $\tilde{c} \in \mathbb{R}^{k^{2m}}$ and $\tilde{u} \in \mathbb{R}^{k^{2m}}$ denote lexicographic representation of $C \in \mathbb{R}^{km \times km}$ and $U \in \mathbb{R}^{km \times km}$, respectively, which leads to $\text{tr}(C^T U) = \tilde{c}^T \tilde{u}$. Thus, the problem in Equation 11 can be recast as a standard quadratic programming problem. To that end, we also need to vectorize the matrix of transportation plan in $p_x = U\kappa^{km}$ and $p_x_r = U^T \kappa^{km}$ such that $p_x = \Omega \tilde{u}$ and $p_x_r = \Lambda \tilde{u}$. Here, $\Omega = [I_{km} | I_{km} | \ldots | I_{km}] \in \mathbb{R}^{km \times k^{2m}}$ is the horizontal concatenation of $km$ identity matrices and $\Lambda = [e_1 | \ldots | e_1 | \ldots | e_{km} | \ldots | e_{km}] \in \mathbb{R}^{km \times k^{2m}}$ is the horizontal concatenation of $km$-dimensional canonical basis, e.g., $e_1 = (1, \ldots, 0)^T$ and $e_{km} = (0, \ldots, 1)^T$.

Consequently, Equation 11 can be rearranged as the following standard quadratic programming problem:

$$J_{WM-VDA}(\tilde{u}) = \frac{1}{2} \tilde{u}^T \left( 2 \Omega^T X^T (B^{-1} + H^T R^{-1} H) X \Omega \right) \tilde{u} + \left( \lambda \tilde{c}^T - 2 (x_b^T B^{-1} + y^T R^{-1} H) X \Omega \right) \tilde{u}.$$

(12)

Thus, the WM-VDA amounts to obtaining the “analysis transportation plan” $\tilde{u}_a$:

$$\tilde{u}_a = \arg\min_{\tilde{u}} J_{WM-VDA}(\tilde{u})$$

subject to $\tilde{u}_i \geq 0$

$$\Lambda \tilde{u} = p_x_r,$$

(13)
which can be efficiently solved through interior-point optimization techniques (Altman and Gondzio, 1999). Finally, the analysis state can be obtained as \( x_a = X\Omega \tilde{u}_a \). It shall be noted that, since the obtained analysis state doesn’t exactly satisfy the constraint of Wasserstein metric, WM-VDA is a weak-constraint DA formulation based on the terminology introduced by Daley (1993).

4. Numerical Experiments and Results

In DA experimentation, we run the forward model under controlled model and observation error, which enables to characterize the effectiveness of the proposed methodology in comparison with the classic 3D-Var approach. To initially examine the performance of the WM-VDA, we focus on two dynamic systems with different levels of sensitivity to their initial conditions including a first order linear and the chaotic Lorenz-63 system. First order dynamical systems have been the cornerstone in developing the Kalman filter (Kalman, 1960) and has been widely used as a test-bed to examine the performance of new filtering techniques (Hazan et al., 2017, 2018). On the other hand, the Lorenz-63 has been the subject of numerous experiments to test the performance of new DA techniques under a chaotic dynamics (Anderson and Anderson, 1999; Miller et al., 1994; Harlim and Hunt, 2007; Van Leeuwen, 2010; Reich, 2012; Goodliff et al., 2015).
a. First Order Linear Dynamics

1) State-space characterization

A first-order discrete time representation of a linear dynamic system in the state-space is presented as follows:

\[ x_{i+1} = M x_i + w_i \]
\[ y_i = H x_i + v_i, \]

(14)

where, \( M \in \mathbb{R}^{m \times m} \) is the (time-invariant) state transition matrix. It is important to note that the system remains stationary if and only if \( \max_i \{ |\gamma_i| \} < 1 \), where \( \{ \gamma_i \}_{i=1}^{m} \) denote the eigenvalues of \( M \).

To examine the effectiveness of the proposed WM-VDA approach in treating systematic biases, we introduce a shift in the system dynamics by assuming that the model and observation errors are drawn from non-zero mean Gaussian distributions.

2) Assimilation set-up and results

Here, we confine our considerations to a 1-D simulation of a linear state space. The initial parameter values were chosen as \( x_0 = 10 \) and \( M = 0.97 \). The expected value (ground truth) of the true state trajectory is generated by solving the model dynamics in Equation 14 with a time step of \( \Delta t = 0.01 \) [t] over a period of \( T = 3 \) [t], in the absence of any model error. For each simulation time step, the model error is drawn from \( \omega_i \sim \mathcal{N}(0, 1.5) \). The observations are obtained by corrupting the truth at assimilation interval \( T_a = 3\Delta t \) using \( v_i \sim \mathcal{N}(0, 25, 0.75) \). To represent the relatively unbiased but highly uncertain a priori reference probability histogram \( p_{x_r} \), 500 samples are drawn at each assimilation intervals from a Gaussian distribution where its mean is located on the ground truth and its variance is set to \( \sigma_x^2 = 3\sigma_b^2 \), where \( \sigma_b^2 = 1.5 \) is the background error covariance. To solve the WM-VDA, the regularization parameter is set to \( \lambda = 5 \). As will be
explained later, this value empirically minimizes the analysis mean squared error (MSE). In order to have a robust conclusion about comparison of the proposed WM-VDA scheme with the classic 3D-Var, the DA experimentation is repeated for 50 independent ensemble simulations.

For all 50 ensembles, the ground truth of the model trajectory, the 2.5 and 97.5 percentiles of the ensemble members and their associated quality metrics, including the bias and unbiased root mean squared error (ubrmse), are shown in Figure 2. As is evident, not only the uncertainty range of the ensemble members (Figure 2a) but also the quality metrics (Figure 2b and c) noticeably improved for the WM-VDA scheme compared to the 3D-Var. In particular, on average, WM-VDA leads to the reduction of bias (ubrmse) from 1.4 to 0.7 (1.6 to 1.3), which is equivalent to 50% (19%) reduction when compared with the 3D-Var.

The sensitivity of the quality metrics is also tested for different range of assimilation intervals \( T_a = \{2\Delta t, 5\Delta t, 10\Delta t, 20\Delta t\} \) for both DA schemes (Figure 3a and b). It is found that the WM-VDA improves the error quality metrics compared to 3D-Var across the chosen range of the assimilation intervals. In particular, for small assimilation intervals of \(2\Delta t\), WM-VDA reduces the bias (ubrmse) from 1 to 0.5 (1.25 to 1), which is equivalent to 50% (20%) reduction when compared to classic 3D-Var. As expected, the ubrmse is reduced less significantly than the bias as the variance of the assimilated reference probability histogram was markedly larger than both observations and background state. As shown, when the assimilation interval grows, the bias and ubrmse in both schemes monotonically increase; however, at different rates. In fact, the bias grows faster in 3D-Var than the WM-VDA while the already small gap between the ubrmse values from the two methods slightly shrinks as the assimilation interval grows. Thus, the analysis state bias in WM-VDA seems to be more robust to increased assimilation intervals than the 3D-Var. This feature needs further investigation as it could be highly desirable for land surface DA, since the satellite overpasses are often available at much longer time intervals than the forecast time steps.
As previously noted, the regularization parameter $\lambda$ plays a significant role in WM-VDA algorithm by making a trade-off between the weighted Euclidean cost and the transportation cost. Recall that larger values of $\lambda$ push or overfits the analysis state towards the \textit{a priori} reference probability histogram by neglecting the information content of background state and observations and thus reduce bias at the expense of an increased spread in uncertainty of the analysis state. On the other hand, smaller values diminish the role of the transportation cost and render it ineffective for bias correction. As previously noted, there is no closed form solution for optimal approximation of this parameter. Here, we focus on determining optimal values for $\lambda$ through cross-validation and trial and error analysis.

Figure 4a demonstrates the evolution of the analysis probability histogram as a function of the regularization parameter $\lambda = \{0.1, 5, 50, 1000\}$, at the first assimilation cycle, for the shown experimental setting in Figure 2. It can be seen that for small values of $\lambda \leq 5$, due to the existing biases, the analysis probability histogram acquires a bimodal distribution as the results approach to those of the 3D-Var scheme. However, as the value of $\lambda$ is increased, the bimodality begins to fade and a more proper geometry of the analysis probability histogram is recovered. At larger values of $\lambda = 1000$, the analysis probability histogram matches perfectly with the \textit{a priori} reference probability histogram as the transportation cost dominates the quadratic cost of the 3D-Var.

Figure 4b shows the variation of the quality metrics as a function of $\lambda$ averaged over 50 independent ensemble simulations. It can be seen that the bias decreases and ubrmse increases for larger values of $\lambda$, yielding a trade-off point where the MSE is minimal. As shown, this minimum MSE is achieved in the range of $\lambda = 5–10$ based on the chosen model parameters and error terms. Clearly, for every problem at hand, this analysis needs to be done offline prior to the implementation of the WM-VDA scheme.
b. Lorenz-63

1) STATE-SPACE CHARACTERIZATION

Lorenz system (Lorenz-63, [Lorenz, 1963]) is a chaotic ordinary differential equation obtained as Fourier truncation of the Rayleigh-Bénard convective flow of fluids in which the three coordinates \(x\), \(y\) and \(z\) represent rate of convective overturn, horizontal and vertical temperature variations, respectively. The system is represented as follows:

\[
\begin{align*}
\frac{dx}{dt} &= -\sigma (x - y) \\
\frac{dy}{dt} &= \rho x - y - xz , \\
\frac{dz}{dt} &= xy - \beta z 
\end{align*}
\] (15)

where, \(\sigma\), \(\rho\) and \(\beta\) are the Prandtl number, a normalized Rayleigh number and a dimensionless wave number respectively. The standard parameter values are \(\sigma = 10\), \(\rho = 28\) and \(\beta = 8/3\) for which the system exhibits a strong chaotic behaviour with maximum Lyapunov exponent value of 0.9. For this selection of parameters, there exist three equilibrium points: the origin that represents the conductive state of no motion and the two attractors located at \((\sqrt{\beta (\rho - 1)}, \sqrt{\beta (\rho - 1)}, \rho - 1)\) and \((-\sqrt{\beta (\rho - 1)}, -\sqrt{\beta (\rho - 1)}, \rho - 1)\), representing a pattern of convective rolls with different directions of rotation.

2) ASSIMILATION SETUP AND RESULTS

In this subsection, we present the results from two different DA settings for the Lorenz system, which differ in terms of characterization of the \textit{a priori} reference probability histogram \(p_{x_r}\). In setup I, it is assumed that \(p_{x_r}\) is available at each assimilation interval while in setup II, it is available over a window of time. The reason for examination of the second experimental setup
is that, in practice, adequate samples to construct the reference histogram may not be available at each assimilation cycle. However, histogram of historical in situ observations is often available over a window of time or space that can be leveraged to reduce the bias in DA systems. An example is the availability of monthly or seasonal probability histogram of gauge measurements of surface soil temperature, moisture or radiosonde data of atmospheric states. To minimize the computational cost of interior-point optimization algorithm, we solved the problem of determining analysis state in each dimension separately by setting $m = 1$ and utilizing marginal probability histograms along each dimension.

In both experiments, to obtain the ground truth of the model trajectory, the Lorenz system is initialized at $(x, y, z) = (3, -3, 12)$ and is integrated using a fourth-order Runge-Kutta (RK4, Runge-Kutta, 1895; Kutta, 1901) approximation with a time step of $\Delta t = 0.01 \, [t]$ over a time period of $T = 0 - 50 \, [t]$ at the end of which the system attains second order stationarity. The observations are obtained at every $10\Delta t$ by corrupting the truth over a simulation period of $T = 0 - 20 \, [t]$ using a Gaussian error $v_i \sim \mathcal{N}(\beta_y, \sigma_y^2 I_3)$ where $\beta_y = 0.15$ and $\sigma_y^2 = 2$. The model is propagated up to $T = 20 \, [t]$ by adding a Gaussian noise $\omega_i \sim \mathcal{N}(\beta_m, \sigma_m^2 I_3)$ to the model state at every $\Delta t$, where $\beta_m = 0.25$ and $\sigma_m^2 = \sqrt{5}$. For setup I, at the assimilation interval of $T_a = 10\Delta t$, 500 samples are drawn from a Gaussian distribution with mean of the ground truth and covariance $\sqrt{3} \sigma_b^2 I_3$ to construct the unbiased a priori reference probability histogram $p_{xr}$. To solve the WM-VDA, the regularization parameter is set to $\lambda = 3$ for each dimension by trial and error. However, in setup II, $p_{xr}$ is constructed through adding a zero-mean Gaussian noise with covariance $\sqrt{3} \sigma_b^2 I_3$ to the ground truth over a period of $T = 0 - 50 \, [t]$. In this case, the $p_{xr}$ has a larger spread than the setup I and provides an unbiased representation of the process over the entire simulation period. In setup II, the regularization parameter is set to $\lambda = (0.02, 0.08, 0.07)$ for the three coordinates. Throughout, in order to have a robust conclusion about comparison of the proposed WM-VDA...
scheme with the 3D-Var, DA experiments are repeated for 50 independent ensemble simulations for both experimental settings.

Figure 5 depicts the trajectory of the ground truth and range showing the 2.5 and 97.5 percentiles of the ensemble members by the 3D-Var and WM-VDA schemes for the first experiment. It is seen that for all three state variables, the range for WM-VDA is narrower than the 3D-Var. On average, bias and ubrmse are reduced by 40–50% and 30–40%, respectively. Clearly the uncertainty of the quality metrics can also be quantified at each time step for all ensemble members. Under a chaotic dynamics, a narrower range for the error statistics signifies more stable solutions to a biased perturbation of the initial conditions. The time evolution of the uncertainty range, between 2.5 and 97.5 percentiles, of the bias and ubrmse is color coded on the phase space of the Lorenz system in Figure 6. The color gradient from dark red to yellow shows that the uncertainty around the computed error statistics gradually shrinks as the DA progresses and passes the duration of spin-up time. It is seen that the width of uncertainty of the quality metrics is noticeably narrower in the WM-VDA than the 3D-Var. In particular, the range reduces by 47.5% and 62.1% for the bias and ubrmse, respectively, demonstrating the advantage of bias-aware WM-VDA over 3D-Var. Table 1 lists the expected values of the bias and ubrmse at the end of both experiments.

To quantify the effects of assimilation intervals on the quality of the DA schemes, experimental setup I was performed for a range of assimilation intervals $T_a = \{2\Delta t, 5\Delta t, 10\Delta t, 20\Delta t\}$. The expected values of the obtained quality metrics are shown in Figure 7. As expected, for both schemes, bias and ubrmse increase as the assimilation interval grows and as shown, the WM-VDA outperforms 3D-Var. The gap between the two approaches, in terms of the bias, remains relatively steady even though we see that over the third dimension, it begins to shrink as the assimilation interval increases. However, it appears that as the assimilation interval grows, the gap between the ubrmse keeps increasing. This might have stemmed from a very high value of Lyapunov exponent.
(0.9) for the Lorenz-63. At such high value, an infinitesimally close trajectories of the state can also deviate significantly as the time progresses. Therefore, for longer assimilation intervals, an analysis state produced by WM-VDA evolves with much less deviation from the true state before the next assimilation cycle and thus exhibit reduced ubrmse than the 3D-Var.

As listed in Table 1 in the second experimental setting, bias in WM-VDA is lower than 3D-Var by 15–50% whereas there is marginal improvement in the ubrmse. The reason is that the relatively unbiased a priori probability histogram is selected over a window of time, with a larger uncertainty ($\sigma_{x_r}^2 = 65 – 82$) along three dimensions than the first experiment ($\sigma_{x_r}^2 = \sqrt{15}$). Thus, its assimilation can only reduce the bias but unable to substantially decrease the ubrmse. However, the results are promising in a sense that even if the spread of a priori reference probability histogram is much larger than the observations and background probability histogram, the WM-VDA can effectively reduce the bias in analysis without hampering the variance. This observation partially verifies the hypothesis that the WM-VDA provides a new way to effectively assimilate highly uncertain but relatively unbiased probability histogram, obtained from the in situ data in a climatological sense.

We also compared the results of setup-II of WM-VDA scheme with the CDF-matching technique implemented on Lorenz-63 system with identical assumptions about error structures as discussed earlier in this section. The CDF-matching scheme is developed without consideration of any a priori reference information; therefore, in its current form, it is limited to bias correction in either model or observation only. For a fair comparison between the methods, with provision of a priori information, we deployed the CDF-matching technique by mapping both observations and model state onto the a priori reference dataset using their respective cumulative histograms. To proceed with CDF-matching, cumulative histogram of the reference dataset is constructed by integrating the model in time from initial state of $(x, y, z) = (3, -3, 12)$ with a time step of $\Delta t = 0.01$ [t] over a period of $T = 0 – 50$ [t] and adding a zero-mean Gaussian noise with covariance $\sqrt{3} \sigma_b^2 I_3$ where,
\( \sigma_b^2 = \sqrt{5} \). The observation cumulative histogram is constructed using observations at an interval of 10\( \Delta t \) over \( T = 0 – 50 \) [t] obtained by corrupting the truth with Gaussian error \( v_i \sim \mathcal{N}(\beta_y^3, \sigma_y^2 I_3) \) where \( \beta_y = 0.15 \) and \( \sigma_y^2 = 2 \). The model cumulative histogram is computed by propagating the model (Equation 15) over \( T = 0 – 50 \) [t] by adding a Gaussian noise \( \omega_i \sim \mathcal{N}(\beta_m^3, \sigma_b^2 I_3) \) to the model state at every \( \Delta t \), where \( \beta_m = 0.25 \) and \( \sigma_b^2 = \sqrt{5} \). At the end of \( T = 50 \) [t], the model state perturbed with zero-mean Gaussian noise of covariance \( \sqrt{3} \sigma_b^2 I_3 \) is considered as the initial condition and DA experimentation for CDF-matching technique is then applied up to the next 2000 time steps. At every assimilation time, observation and model are first mapped onto the reference dataset using piece-wise linear CDF-matching to remove bias and then the 3D-Var method was utilized to obtain the analysis state using bias removed model state and observation. It is to note that, in comparison to the WM-VDA, CDF-matching is equipped with more information fed in the form of model and observation cumulative mass functions.

It is seen that over 50 independent ensemble simulations, averaged over all three dimensions, CDF-matching reduces the bias by 26% compared to 28% by WM-VDA whereas ubrmse increased by 8.6% compared to overall decrease in ubrmse of 3.3% by WM-VDA. The difference exists primarily due to the way CDF-matching completely maps the biased source of information onto the relatively unbiased one. Here, since the reference dataset has higher uncertainty in terms of variance, CDF-matching improves bias compared to 3D-Var at the expense of increase in ubrmse. Therefore, if the a priori information has uncertainty, as can be expected in real condition, WM-VDA can be more effective than a statistic CDF-matching technique in bias correction and also provide a fertile ground to conduct DA experiments in probability domain with reduced uncertainty.
5. Summary and Concluding Remarks

In this study, we discussed the concept of Wasserstein metric (WM) regularization of the variational data assimilation (VDA) techniques, referred to as WM-VDA. In particular, the cost of the classic VDA problem is equipped with a transportation cost that penalizes the mismatch between the probability histograms of the analysis state and a relatively unbiased \textit{a priori} reference data, which can be obtained from \textit{in-situ} observations over a spatial or temporal window. The presented WM-VDA approach does not need any \textit{a priori} knowledge on the sign of the bias and information about its origin from either model and/or observations and automatically retrieve it from the \textit{a priori} reference data. We examined the application of WM-VDA for bias removal in a simple first-order linear dynamics and the chaotic Lorenz-63 system (Lorenz, 1963). The results demonstrated that WM-VDA can reduce and control propagation of bias under chaotic dynamics. Due to the intrinsic property of the Wasserstein metric to allow natural morphing between the probability histograms (Kolouri \textit{et al.}, 2017), a proper value of the regularization parameter alleviates any unrealistic bimodality in the shape of the analysis histogram due to potential biases. Initial comparisons with classic CDF matching also demonstrated improved performance, although future research is needed for a comprehensive comparison with other existing methods to fully characterize the advantages and disadvantages of the proposed approach.

As it is well understood, in the absence of bias, the classic VDA leads to the lowest possible analysis mean squared error and meets the known Cramér-Rao lower bound (CRLB, Rao \textit{et al.}, 1973; Cramér, 1999). However, in the presence of systematic biases such a lower bound cannot be met. Even though, we empirically demonstrated that under biased assimilation scenarios, the WM-VDA shows improved performance than the 3D-VAR, future theoretical studies are needed to characterize a closed form expression for such an improvement. It is to note that, since WM-VDA
doesn’t need to specifically attribute the bias proportionally to either model or observations, it cannot identify the origin of bias. If such information is needed, future research might be devoted to relate the amount of transported probability masses to the bias of the background state and observations.

One of the main challenges of WM-VDA is its high computational complexity. The interior-point optimization algorithm has computational complexity $O(k^3 \log k)$ where, $k$ is the number of elements in the support set. This hampers applications of the WM-VDA to large-scale geophysical DA problems. Recent advances in tomographic approximation of the Wasserstein distance (Kolouri et al., 2018) through slicing the metric via a finite number of 1-D Radon projections can significantly reduce its computational cost for high-dimensions problems and can be a possible direction for future research in geophysical data assimilation. Moreover, characterization of the regularization parameter through cross validation could be computationally intensive, especially for large-scale Earth system models and new research efforts are required to address this challenge.

As demonstrated empirically, the WM-VDA scheme shows robustness to increased assimilation intervals, which can pave new ways for improving bias-aware satellite VDA systems. Of particular interest is satellite soil moisture or precipitation DA (Lin et al., 2015, 2017a) as the land and weather models are often biased (Reichle et al., 2004; Lin et al., 2017b). To that end, future research is required to understand how WM-VDA can be integrated into land-weather models for assimilation of in-situ data in probability domain. A promising area is to incorporate the Wasserstein metric for bias correction in ensemble-based iterative methods and hybrid variational-ensemble DA techniques. In particular, the theory of optimal mass transport seems to provide a concrete ground to solve the problem of filter degeneracy in particle filters (Snyder et al., 2008; Van Leeuwen, 2010; Reich and Cotter, 2015; Van Leeuwen et al., 2019), where the observations
and background state do not have overlapping supports and thus the weights cannot be trivially updated using the likelihood function.

Acknowledgements

The first and second authors acknowledge the grant from the National Aeronautics and Space Administration (NASA) Terrestrial Hydrology Program (THP, 80NSSC18K1528) and the New (Early Career) Investigator Program (NIP, 80NSSC18K0742).

References

Altman, A., and J. Gondzio (1999), Regularized symmetric indefinite systems in interior point methods for linear and quadratic optimization, *Optimization Methods and Software, 11*(1-4), 275–302.

Anderson, J. L., and S. L. Anderson (1999), A monte carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts, *Monthly Weather Review, 127*(12), 2741–2758.

Asadi, N., K. A. Scott, and D. A. Clausi (2019), Data fusion and data assimilation of ice thickness observations using a regularisation framework, *Tellus A: Dynamic Meteorology and Oceanography*, pp. 1–20.

Asch, M., M. Bocquet, and M. Nodet (2016), *Data assimilation: methods, algorithms, and applications*, vol. 11, SIAM.

Auligné, T., A. McNally, and D. Dee (2007), Adaptive bias correction for satellite data in a numerical weather prediction system, *Quarterly Journal of the Royal Meteorological Society: A jour-
nal of the atmospheric sciences, applied meteorology and physical oceanography, 133(624), 631–642.

Benamou, J.-D., G. Carlier, M. Cuturi, L. Nenna, and G. Peyré (2015), Iterative bregman projections for regularized transportation problems, SIAM Journal on Scientific Computing, 37(2), A1111–A1138.

Brenier, Y. (1991), Polar factorization and monotone rearrangement of vector-valued functions, Communications on pure and applied mathematics, 44(4), 375–417.

Charney, J. G. (1951), Dynamic forecasting by numerical process, in Compendium of meteorology, pp. 470–482, Springer.

Chen, Y., T. T. Georgiou, and A. Tannenbaum (2017), Matrix optimal mass transport: a quantum mechanical approach, IEEE Transactions on Automatic Control, 63(8), 2612–2619.

Chen, Y., T. T. Georgiou, and A. Tannenbaum (2018a), Optimal transport for gaussian mixture models, IEEE Access, 7, 6269–6278.

Chen, Y., T. T. Georgiou, and A. Tannenbaum (2018b), Wasserstein geometry of quantum states and optimal transport of matrix-valued measures, in Emerging Applications of Control and Systems Theory, pp. 139–150, Springer.

Chepurin, G. A., J. A. Carton, and D. Dee (2005), Forecast model bias correction in ocean data assimilation, Monthly weather review, 133(5), 1328–1342.

Courtier, P., J.-N. Thépaut, and A. Hollingsworth (1994), A strategy for operational implementation of 4d-var, using an incremental approach, Quarterly Journal of the Royal Meteorological Society, 120(519), 1367–1387.
Cramér, H. (1999), *Mathematical methods of statistics*, vol. 9, Princeton university press.

Crow, W. T., R. D. Koster, R. H. Reichle, and H. O. Sharif (2005), Relevance of time-varying and time-invariant retrieval error sources on the utility of spaceborne soil moisture products, *Geophysical Research Letters*, 32(24).

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

De Lannoy, G. J. M., P. de Rosnay, and R. H. Reichle (2016), Soil moisture data assimilation, *Handbook of Hydrometeorological Ensemble Forecasting*, pp. 1–43.

Dee, D. P. (2003), Detection and correction of model bias during data assimilation, *Meteorological Training Course Lecture Series (ECMWF)*.

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

Dee, D. P., and A. M. Da Silva (1998), Data assimilation in the presence of forecast bias, *Quarterly Journal of the Royal Meteorological Society*, 124(545), 269–295.

Dee, D. P., and S. Uppala (2009), Variational bias correction of satellite radiance data in the era-interim reanalysis, *Quarterly Journal of the Royal Meteorological Society*, 135(644), 1830–1841.

Drécourt, J.-P., H. Madsen, and D. Rosbjerg (2006), Bias aware kalman filters: Comparison and improvements, *Advances in Water Resources*, 29(5), 707–718.

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(9), 5944–5963.
Ebtehaj, A. M., M. Zupanski, G. Lerman, and E. Foufoula-Georgiou (2014), Variational data assimilation via sparse regularisation, *Tellus A: Dynamic Meteorology and Oceanography*, 66(1), 21,789.

Engquist, B., and B. D. Froese (2013), Application of the wasserstein metric to seismic signals, *arXiv preprint arXiv:1311.4581*.

Feyeux, N., A. Vidard, and M. Nodet (2018), Optimal transport for variational data assimilation, *Nonlinear Processes in Geophysics*, 25(1), 55–66.

Freitag, M. A., N. K. Nichols, and C. J. Budd (2010), L1-regularisation for ill-posed problems in variational data assimilation, *Pamm*, 10(1), 665–668.

Gangbo, W., and R. J. McCann (1996), The geometry of optimal transportation, *Acta Mathematica*, 177(2), 113–161.

Goodliff, M., J. Amezcua, and P. J. Van Leeuwen (2015), Comparing hybrid data assimilation methods on the lorenz 1963 model with increasing non-linearity, *Tellus A: Dynamic Meteorology and Oceanography*, 67(1), 26,928.

Hamilton, F., T. Berry, and T. Sauer (2019), Correcting observation model error in data assimilation, *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(5), 053,102.

Harlim, J., and B. R. Hunt (2007), A non-gaussian ensemble filter for assimilating infrequent noisy observations, *Tellus A: Dynamic Meteorology and Oceanography*, 59(2), 225–237.

Hazan, E., K. Singh, and C. Zhang (2017), Learning linear dynamical systems via spectral filtering, in *Advances in Neural Information Processing Systems*, pp. 6702–6712.

Hazan, E., H. Lee, K. Singh, C. Zhang, and Y. Zhang (2018), Spectral filtering for general linear dynamical systems, in *Advances in Neural Information Processing Systems*, pp. 4634–4643.
Jin, J., H. X. Lin, A. Segers, Y. Xie, and A. Heemink (2019), Machine learning for observation bias correction with application to dust storm data assimilation, *Atmospheric Chemistry and Physics*, 19(15), 10,009–10,026.

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

Kalnay, E. (2003), *Atmospheric modeling, data assimilation and predictability*, Cambridge university press.

Kalnay, E., H. Li, T. Miyoshi, S.-C. Yang, and J. Ballabrera-Poy (2007), 4-d-var or ensemble kalman filter?, *Tellus A: Dynamic Meteorology and Oceanography*, 59(5), 758–773.

Kantorovich, L. V. (1942), On the translocation of masses, in *Dokl. Akad. Nauk. USSR (NS)*, vol. 37, pp. 199–201.

Kolouri, S., S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde (2017), Optimal mass transport: Signal processing and machine-learning applications, *IEEE signal processing magazine*, 34(4), 43–59.

Kolouri, S., G. K. Rohde, and H. Hoffmann (2018), Sliced wasserstein distance for learning gaussian mixture models, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3427–3436.

Kumar, S. V., R. H. Reichle, R. D. Koster, W. T. Crow, and C. D. Peters-Lidard (2009), Role of subsurface physics in the assimilation of surface soil moisture observations, *Journal of hydrometeorology*, 10(6), 1534–1547.
Kumar, S. V., R. H. Reichle, K. W. Harrison, C. D. Peters-Lidard, S. Yatheendradas, and J. A. Santanello (2012), A comparison of methods for a priori bias correction in soil moisture data assimilation, *Water Resources Research, 48*(3).

Kutta, W. (1901), Beitrag zur naherungsweisen integration totaler differentialgleichungen, *Z. Math. Phys., 46*, 435–453.

Law, K. J., and A. M. Stuart (2012), Evaluating data assimilation algorithms, *Monthly Weather Review, 140*(11), 3757–3782.

Leith, C. (1993), Numerical models of weather and climate, *Plasma physics and controlled fusion, 35*(8), 919.

Li, P., Q. Wang, and L. Zhang (2013), A novel earth mover’s distance methodology for image matching with gaussian mixture models, in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1689–1696.

Lin, L.-F., A. M. Ebtehaj, R. L. Bras, A. N. Flores, and J. Wang (2015), Dynamical precipitation downscaling for hydrologic applications using wrf 4d-var data assimilation: Implications for gpm era, *Journal of Hydrometeorology, 16*(2), 811–829.

Lin, L.-F., A. M. Ebtehaj, A. N. Flores, S. Bastola, and R. L. Bras (2017a), Combined assimilation of satellite precipitation and soil moisture: A case study using trmm and smos data, *Monthly Weather Review, 145*(12), 4997–5014.

Lin, L.-F., A. M. Ebtehaj, J. Wang, and R. L. Bras (2017b), Soil moisture background error covariance and data assimilation in a coupled land-atmosphere model, *Water Resources Research, 53*(2), 1309–1335.
Liu, Q., R. H. Reichle, R. Bindlish, M. H. Cosh, W. T. Crow, R. de Jeu, G. J. De Lannoy, G. J. Huffman, and T. J. Jackson (2018), The contributions of precipitation and soil moisture observations to the skill of soil moisture estimates in a land data assimilation system, *Journal of Hydrometeorology*, 19(2).

Lorenc, A. C. (1986), Analysis methods for numerical weather prediction, *Quarterly Journal of the Royal Meteorological Society*, 112(474), 1177–1194.

Lorenc, A. C., N. E. Bowler, A. M. Clayton, S. R. Pring, and D. Fairbairn (2015), Comparison of hybrid-4denvar and hybrid-4dvar data assimilation methods for global nwp, *Monthly Weather Review*, 143(1), 212–229.

Lorenz, E. N. (1963), Deterministic nonperiodic flow, *Journal of the atmospheric sciences*, 20(2), 130–141.

Miller, R. N., M. Ghil, and F. Gauthiez (1994), Advanced data assimilation in strongly nonlinear dynamical systems, *Journal of the atmospheric sciences*, 51(8), 1037–1056.

Monge, G. (1781), Mémoire sur la théorie des déblais et des remblais, *Histoire de l'Académie Royale des Sciences de Paris*.

Motamed, M., and D. Appelo (2018), Wasserstein metric-driven bayesian inversion with application to wave propagation problems, *arXiv preprint arXiv:1807.09682*.

Ning, L., F. P. Carli, A. M. Ebtehaj, E. Foufoula-Georgiou, and T. T. Georgiou (2014), Coping with model error in variational data assimilation using optimal mass transport, *Water Resources Research*, 50(7), 5817–5830.
Pauwels, V. R., G. J. De Lannoy, H.-J. Hendricks Franssen, and H. Vereecken (2013), Simultaneous estimation of model state variables and observation and forecast biases using a two-stage hybrid kalman filter, *Hydrology and Earth System Sciences*, 17(9), 3499–3521.

Peyré, G., M. Cuturi, et al. (2019), Computational optimal transport, *Foundations and Trends® in Machine Learning*, 11(5-6), 355–607.

Rabier, F. (2005), Overview of global data assimilation developments in numerical weather-prediction centres, *Quarterly Journal of the Royal Meteorological Society*, 131(613), 3215–3233.

Radakovich, J. D., P. R. Houser, A. da Silva, and M. G. Bosilovich (2001), Results from global land-surface data assimilation methods, in *AGU Spring Meeting Abstracts*.

Rao, C. R., C. R. Rao, M. Statistiker, C. R. Rao, and C. R. Rao (1973), *Linear statistical inference and its applications*, vol. 2, Wiley New York.

Reich, S. (2012), A gaussian-mixture ensemble transform filter, *Quarterly Journal of the Royal Meteorological Society*, 138(662), 222–233.

Reich, S., and C. Cotter (2015), *Probabilistic forecasting and Bayesian data assimilation*, Cambridge University Press.

Reichle, R. H., and R. D. Koster (2004), Bias reduction in short records of satellite soil moisture, *Geophysical Research Letters*, 31(19).

Reichle, R. H., R. D. Koster, J. Dong, and A. A. Berg (2004), Global soil moisture from satellite observations, land surface models, and ground data: Implications for data assimilation, *Journal of Hydrometeorology*, 5(3), 430–442.
Reichle, R. H., R. D. Koster, P. Liu, S. P. Mahanama, E. G. Njoku, and M. Owe (2007), Comparison and assimilation of global soil moisture retrievals from the advanced microwave scanning radiometer for the earth observing system (amsr-e) and the scanning multichannel microwave radiometer (smmr), *Journal of Geophysical Research: Atmospheres, 112*(D9).

Reichle, R. H., S. V. Kumar, S. P. Mahanama, R. D. Koster, and Q. Liu (2010), Assimilation of satellite-derived skin temperature observations into land surface models, *Journal of Hydrometeorology, 11*(5), 1103–1122.

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.

Runge, C. (1895), Über die numerische auflösung von differentialgleichungen, *Mathematische Annalen, 46*(2), 167–178.

Santambrogio, F. (2015), Optimal transport for applied mathematicians, *Birkäuser, NY*, 55, 58–63.

Snyder, C., T. Bengtsson, P. Bickel, and J. Anderson (2008), Obstacles to high-dimensional particle filtering, *Monthly Weather Review, 136*(12), 4629–4640.

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.

Van Leeuwen, P. J., H. R. Künsch, L. Nerger, R. Potthast, and S. Reich (2019), Particle filters for high-dimensional geoscience applications: A review, *Quarterly Journal of the Royal Meteorological Society*.

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

Villani, C. (2008), *Optimal transport: old and new*, vol. 338, Springer Science & Business Media.
Villarini, G., P. V. Mandapaka, W. F. Krajewski, and R. J. Moore (2008), Rainfall and sampling uncertainties: A rain gauge perspective, *Journal of Geophysical Research: Atmospheres, 113*(D11).

Wahba, G., and J. Wendelberger (1980), Some new mathematical methods for variational objective analysis using splines and cross validation, *Monthly weather review, 108*(8), 1122–1143.

Wang, X., D. M. Barker, C. Snyder, and T. M. Hamill (2008), A hybrid etkf–3dvar data assimilation scheme for the wrf model. part i: Observing system simulation experiment, *Monthly Weather Review, 136*(12), 5116–5131.

Zhu, Y., J. Derber, A. Collard, D. Dee, R. Treadon, G. Gayno, and J. A. Jung (2014), Enhanced radiance bias correction in the national centers for environmental prediction’s gridpoint statistical interpolation data assimilation system, *Quarterly Journal of the Royal Meteorological Society, 140*(682), 1479–1492.

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

Table 1. Expected values of the bias and ubrmse in 3D-Var and WM-VDA for the experimental setup I and II. Values shown inside of the parentheses are percent reduction of the error metrics due to implementation of the WM-VDA compared to the 3D-Var.
TABLE 1. Expected values of the bias and ubrmse in 3D-Var and WM-VDA for the experimental setup I and II. Values shown inside of the parentheses are percent reduction of the error metrics due to implementation of the WM-VDA compared to the 3D-Var.

| Methods       | bias x | y | z | ubrmse x | y | z |
|---------------|--------|---|---|----------|---|---|
| 3D-Var        | 2.65   | 2.17| 2.22| 3.92     | 5.71| 5.36 |
| setup I: WM-VDA | 1.31 (50.6) | 1.28 (41.0) | 1.34 (39.6) | 2.83 (27.8) | 3.39 (40.6) | 3.29 (38.6) |
| setup II: WM-VDA | 2.22 (16.2) | 1.67 (23.0) | 1.20 (45.9) | 3.88 (1.0) | 5.42 (5.1) | 5.19 (3.2) |
LIST OF FIGURES

Fig. 1.  (a) The optimal mass transport problem between source $p_x$ and two target probability histograms $p_z$ and $p_{\hat{z}}$ that are apart from each other only by a first order shift, (b) an example of the Monge formulation of the transportation problem where each source probability mass is assign to only a single target location and (c) the Kantorovich formulation of the problem allowing mass splitting across several target locations. Here, the Euclidean distance ($d_2^2$) between the source and target probability histograms remains insensitive to the shift in position, while the 2-Wasserstein distance ($W_2^2$) increases monotonically as the shift between the central positions of the probability histograms increases.

Fig. 2.  Time evolution of the uncertainty range (colored shaded region) for the 3D-Var (green) and WM-VDA (orange)—representing (a) the 2.5 and 97.5 percentiles of 50 ensemble members of the analysis state $x_a$ and the 95% confidence bound for its (b) bias and (c) ubrmse. The true state ($x_{tr}$) and observations (Obs.) for an independent simulation are shown in the left panel as well.

Fig. 3.  Variation of (a) bias and (b) ubrmse for the 3D-Var and WM-VDA as a function of assimilation interval.

Fig. 4.  (a) Transition of the analysis state probability histogram at first assimilation cycle and (b) variation of bias, ubrmse and mean squared error (MSE) of the analysis error as a function of $\lambda$, averaged over 50 independent ensemble simulations.

Fig. 5.  Time evolution of the true state (solid lines) of the Lorenz-63 analysis states, observations (Obs.) for one ensemble member and 2.5 and 97.5 percentiles (shaded areas) of the ensemble members for the 3D-Var (a, c, e) and the WM-VDA (b, d, f). The results are obtained for 50 independent ensemble members based on the experimental setup I.

Fig. 6.  Time evolution of the 95% confidence bound of the (a,b) bias and (c,d) ubrmse along the phase space of the Lorenz system for the 3D-Var and WM-VDA. The uncertainty range is obtained from 50 independent ensemble members for the first experimental setup.

Fig. 7.  Changes of (a) bias and (b) ubrmse as a function of assimilation interval for 3D-Var and WM-VDA in experimental setup I, where $(x, y, z)$ represents the three dimensions of the Lorenz system.
FIG. 1. (a) The optimal mass transport problem between source \( p_x \) and two target probability histograms \( p_{\hat{z}} \) and \( p_{\hat{z}} \) that are apart from each other only by a first order shift, (b) an example of the Monge formulation of the transportation problem where each source probability mass is assigned to only a single target location and (c) the Kantorovich formulation of the problem allowing mass splitting across several target locations. Here, the Euclidean distance (\( d_2 \)) between the source and target probability histograms remains insensitive to the shift in position, while the 2-Wasserstein distance (\( W_2 \)) increases monotonically as the shift between the central positions of the probability histograms increases.
FIG. 2. Time evolution of the uncertainty range (colored shaded region) for the 3D-Var (green) and WM-VDA (orange)—representing (a) the 2.5 and 97.5 percentiles of 50 ensemble members of the analysis state $x_a$ and the 95% confidence bound for its (b) bias and (c) ubrmse. The true state ($x_{tr}$) and observations (Obs.) for an independent simulation are shown in the left panel as well.
Fig. 3. Variation of (a) bias and (b) ubrmse for the 3D-Var and WM-VDA as a function of assimilation interval.
FIG. 4. (a) Transition of the analysis state probability histogram at first assimilation cycle and (b) variation of bias, ubrmse and mean squared error (MSE) of the analysis error as a function of $\lambda$, averaged over 50 independent ensemble simulations.
Fig. 5. Time evolution of the true state (solid lines) of the Lorenz-63 analysis states, observations (Obs.) for one ensemble member and 2.5 and 97.5 percentiles (shaded areas) of the ensemble members for the 3D-Var (a, c, e) and the WM-VDA (b, d, f). The results are obtained for 50 independent ensemble members based on the experimental setup I.
FIG. 6. Time evolution of the 95% confidence bound of the (a,b) bias and (c,d) ubrmse along the phase space of the Lorenz system for the 3D-Var and WM-VDA. The uncertainty range is obtained from 50 independent ensemble members for the first experimental setup.
Fig. 7. Changes of (a) bias and (b) ubrmse as a function of assimilation interval for 3D-Var and WM-VDA in experimental setup I, where \((x, y, z)\) represents the three dimensions of the Lorenz system.