Received: 24 October 2019 | Revised: 14 February 2020 | Accepted: 18 March 2020 | Published on: 29 April 2020

DOI: 10.1002/qj.3794

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

Regularized variational data assimilation for bias treatment using the Wasserstein metric

Sagar K. Tamang[1] | Ardeshir Ebtehaj[1] | Dongmian Zou[2] | Gilad Lerman[2]

[1] Department of Civil, Environmental and Geo-Engineering and Saint Anthony Falls Laboratory, University of Minnesota–Twin Cities, Twin Cities, Minnesota

[2] School of Mathematics, University of Minnesota–Twin Cities, Twin Cities, Minnesota

Correspondence
Sagar K. Tamang, Department of Civil, Environmental and Geo-Engineering and Saint Anthony Falls Laboratory, University of Minnesota–Twin Cities, Twin Cities, Minnesota.
Email: taman011@umn.edu

Funding information
National Aeronautics and Space Administration (NASA) Terrestrial Hydrology Program (THP, 80NSSC18K1528), the New (Early Career) Investigator Program (NIP, 80NSSC18K0742) and National Science Foundation (NSF,1830418).

Abstract

This article 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) treats systematic biases of unknown magnitude and sign dynamically in both model and observations, through assimilation of the reference data in the probability domain, and can recover the probability histogram of the analysis state fully. The performance of WM-VDA is compared with the classic three-dimensional VDA (3D-Var) scheme for 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.

KEYWORDS

bias treatment, chaotic systems, optimal mass transport, regularization, variational data assimilation, Wasserstein distance

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 shows low forecast skill 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 ESM initial conditions, thus reducing their forecast uncertainties (Leith, 1993; Kalnay, 2003). DA methods often involve iterative cycles, in which the *observations* are integrated optimally 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, which aim to combine and take advantage of the unique benefits of the two DA classes (Wang *et al.*, 2008; Lorenc *et al.*, 2015), are also being developed.

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, due mainly to underrepresentation of the governing laws of physics and erroneous boundary conditions. Observation bias also exists, due largely 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 is known, which is often not 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 closer 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 fall mainly under two general categories: (a) dynamic bias-aware, and (b) rescaling 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). Dynamic bias-aware schemes make prior assumptions about the nature of the bias and attribute it to either model or observations, which may not be realistic, as both models and observations suffer from systematic errors. Early attempts to treat model biases dynamically 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 a bias-aware Kalman filter for coloured noise (Drécourt *et al.*, 2006) and a weak constrained four-dimensional VDA (4D-Var: Zupanski, 1997) have also been proposed to account for nonzero mean model errors. At the same time, another class of dynamic observational bias correction techniques exists that relies on variants of the variational bias-correction (VarBC) method, which makes an a priori estimate of bias and updates it dynamically using innovation information (Auligné *et al.*, 2007; Dee and Uppala, 2009; Zhu *et al.*, 2014). Apart from VarBC, more recently a new approach has been proposed to treat observation biases by iterative updates of the observation operator (Hamilton *et al.*, 2019). A body of research also has been devoted to treating model and observation biases simultaneously using a multistage 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 have not yet been tested for effective bias correction in chaotic systems. More importantly, the developed schemes focus largely on retrieving an unbiased expected value of the forecast and remain limited to characterization of the second-order forecast uncertainty.

The rescaling techniques do not make any explicit assumptions about the relative accuracy of the model and observation system (Reichle and Koster, 2004; Crow *et al.*, 2005; Reichle *et al.*, 2007; 2010; Kumar *et al.*, 2009; Liu *et al.*, 2018). This family of methods often involves mapping the observations on to 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 assumes implicitly that model forecasts are unbiased and partly ignores the information content of observations. For example, if our observations are less biased than the model outputs, this approach basically fails to remove the bias effectively. Furthermore, it is a static scheme and no formal way exists to extend the CDF-matching scheme to account dynamically 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 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 concrete ground on which to compare probability measures (Brenier, 1991; Gangbo and McCann, 1996; Benamou *et al.*, 2015; Chen *et al.*, 2017; 2018a; 2018b). 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 a metric in the VDA cost function can reduce analysis biases. The rationale is that the work done during displacement of the probability masses is a function of not only the shape of the probability histograms but also the difference between their central positions, as described in section 2.3. 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 the 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 replace the quadratic costs in the classic VDA fully 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 article.

This article presents a new VDA approach through regularizing the classic VDA problem with the 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 in both 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 the 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, 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, which represents the nonlinear dynamics of a convective circulation in a shallow fluid layer. The results demonstrate that the approach presented is capable of preserving the geometric shape of the distribution of the analysis state when both the background state and the observations are systematically biased and extending the forecast skills by controlling the propagation of bias in the phase space of a highly chaotic system.

The article is organized as follows. Section 2 discusses the concept of classic VDA, focusing on 3D-Var. In this section, a summary of 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 WM-VDA on a first-order linear system and the nonlinear Lorenz-63 dynamic system. The results are interpreted and compared with the 3D-Var and CDF-matching techniques. A summary and concluding remarks are presented in section 5.

2 | METHODOLOGY

2.1 | Notation

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 $\mathbf{X} \in \mathbb{R}^{m\times n}$, $\mathbb{1}_m$ 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 $\mathbf{p}_x = (p_{x_1}, \dots, p_{x_k})^{\mathrm{T}}$ supported on k points $x_1, \dots, x_k$, such that $\mathbf{x} = \sum_k p_{x_k} \delta_{x_k}$, where δ_{x_k} is the Dirac function at x_k and $(\cdot)^{\mathrm{T}}$ denotes the transposition operator. For the state $\mathbf{x} \in \mathbb{R}^m$, this linear expectation operator is represented as $\mathbf{x} = \mathbf{X}\,\mathbf{p}_x$, where the support point and associated probability of occurrences are properly concatenated in $\mathbf{X} \in \mathbb{R}^{m\times k^m}$ and $\mathbf{p}_x \in \mathbf{R}^{k^m}$, respectively. $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes that $\mathbf{x}$ is drawn from a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance Σ, and the square of the weighted ℓ_2-norm of $\mathbf{x}$ is represented as $\|\mathbf{x}\|^2_{\mathbf{B}^{-1}} = \mathbf{x}^{\mathrm{T}}\mathbf{B}^{-1}\mathbf{x}$, where $\mathbf{B}$ is a positive-definite matrix.

2.2 | Classic 3D-Var

In three-dimensional VDA (3D-Var: 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 $\mathbf{x}_{i+1} \in \mathbb{R}^m$ with the following stochastic dynamics:

$$\mathbf{x}_{i+1} = \mathcal{M}(\mathbf{x}_i) + \boldsymbol{\omega}_i, \tag{1}$$

where $\mathcal{M} : \mathbb{R}^m \rightarrow \mathbb{R}^m$ is the nonlinear model operator evolving the state from $\mathbf{x}_i$ to $\mathbf{x}_{i+1}$ and $\boldsymbol{\omega}_i \sim \mathcal{N}(0, \mathbf{B}) \in \mathbb{R}^m$ is the model error, and the expected background state at the $(i + 1)$th time step is $\mathbf{x}_{\mathrm{b}} = \mathcal{M}(\mathbf{x}_i)$.

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

$$\mathbf{y}_i = \mathcal{H}(\mathbf{x}_i) + \boldsymbol{\nu}_i, \tag{2}$$

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

In a 3D-Var setting, the cost function is comprised of two weighted Euclidean distances of the unknown true state from the background state $\mathbf{x}_{\mathrm{b}}$ and the observation $\mathbf{y}$, such that

$$\mathcal{J}_{3\mathrm{D}}(\mathbf{x}_0) = \|\mathbf{x}_0 - \mathbf{x}_{\mathrm{b}}\|^2_{\mathbf{B}^{-1}} + \|\mathbf{y} - \mathcal{H}(\mathbf{x}_0)\|^2_{\mathbf{R}^{-1}}\,. \tag{3}$$

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

$$\mathbf{x}_{\rm a} = \underset{\mathbf{x}_0}{\mathrm{argmin}} \mathcal{J}_{\rm 3D}(\mathbf{x}_0)$$
$$= \underset{\mathbf{x}_0}{\mathrm{argmin}} \{ \|\mathbf{x}_0 - \mathbf{x}_{\rm b}\|^2_{\mathbf{B}^{-1}} + \|\mathbf{y} - \mathbf{H}\,\mathbf{x}_0\|^2_{\mathbf{R}^{-1}} \}. \qquad (4)$$

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 $\mathbf{x}_{\rm a}$ is obtained as

$$\mathbf{x}_{\rm a} = (\mathbf{H}^{\rm T}\mathbf{R}^{-1}\mathbf{H} + \mathbf{B}^{-1})^{-1}(\mathbf{H}^{\rm T}\mathbf{R}^{-1}\mathbf{y} + \mathbf{B}^{-1}\mathbf{x}_{\rm b}), \qquad (5)$$

where the analysis-error covariance $\mathbf{P}_{\rm a}$ is the inverse of the Hessian of Equation 3 (Daley, 1993):

$$\mathbf{P}_{\rm a} = (\mathbf{H}^{\rm T}\mathbf{R}^{-1}\mathbf{H} + \mathbf{B}^{-1})^{-1}. \qquad (6)$$

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

2.3 | 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 have provided fertile ground for comparison of probability measures (Brenier, 1991; Villani, 2003) and have been studied extensively and applied widely in the fields of signal processing (Kolouri *et al.*, 2017; Motamed and Appelo, 2019), image retrieval (Rubner *et al.*, 2000; Li *et al.*, 2013), and analyzing misfit in seismic signals (Engquist and Froese, 2013).

Let $\mathbf{p}_x = (p_{x_1}, \ldots, p_{x_k})^{\rm T} \in \mathbb{R}^k$ and $\mathbf{p}_z = (p_{z_1}, \ldots, p_{z_l})^{\rm T} \in \mathbb{R}^l$ 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 the 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 the source probability histogram to a "single" point z_j on the target probability histogram, where $i = 1 \ldots, k$ and $j = 1 \ldots, l$, such that the total cost of transportation is minimized:

$$\underset{T(\cdot)}{\text{minimize}} \quad \sum_i c\,(x_i, T(x_i))$$
$$\text{subject to} \quad p_{z_j} = \sum_{i\,:\,T(x_i)=z_j} p_{x_i}, \qquad (7)$$

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, the Monge formulation of the OMT problem is nonconvex and becomes a combinatorial Non-deterministic Polynomial time hard problem, for which the transport map $T(\cdot)$ does not exist—when the 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 the Monge and Kantorovich formulations of the OMT is shown in Figure 1. Let us assume that $\mathbf{U} \in \mathbb{R}^{k \times l}$ 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, $\mathbf{C} \in \mathbb{R}^{k \times l}$ 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:

$$\mathcal{W}_p(\mathbf{p}_x, \mathbf{p}_z) := (\underset{u_{ij}}{\text{minimize}} \sum_{i,j} c_{ij}u_{ij})^{1/p}$$
$$= (\underset{\mathbf{U}}{\text{minimize}} \ \mathrm{tr}(\mathbf{C}^{\rm T}\mathbf{U}))^{1/p}$$
$$\text{subject to } u_{ij} \geq 0,$$
$$\mathbf{U}\mathbb{1}_l = \mathbf{p}_x,$$
$$\mathbf{U}^{\rm T}\mathbb{1}_k = \mathbf{p}_z. \qquad (8)$$

The first constraint ensures that the transported probability masses are nonnegative 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 required 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 the shift between central positions of the histograms (Ning *et al.*, 2014), enabling it to penalize

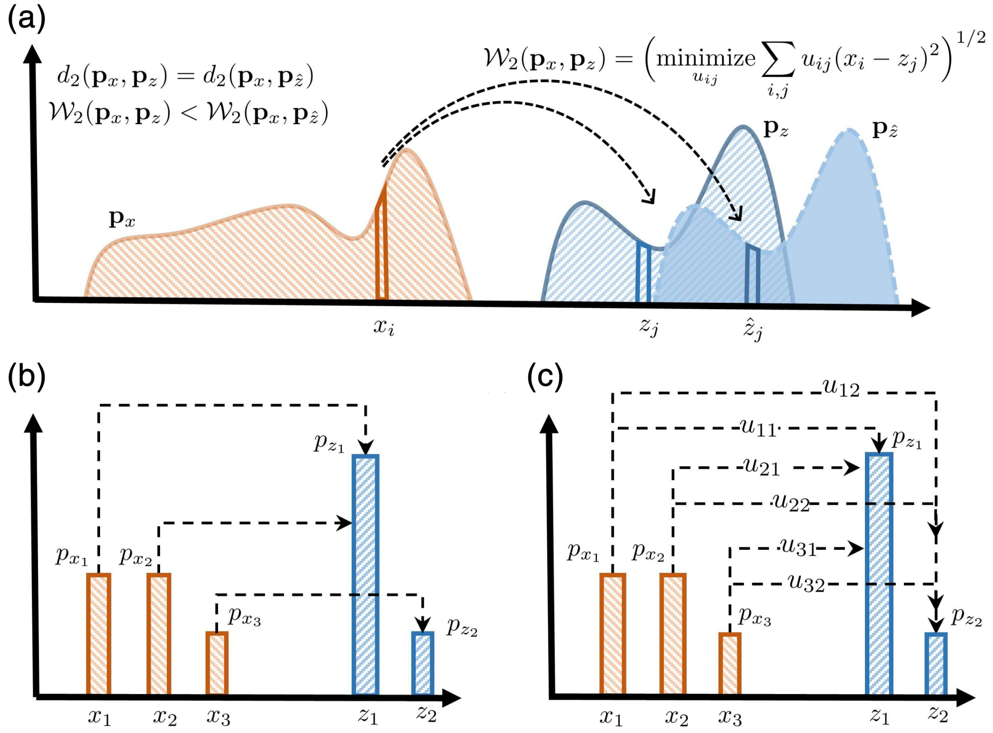
(a)
$d_2(\mathbf{p}_x, \mathbf{p}_z) = d_2(\mathbf{p}_x, \mathbf{p}_{\hat{z}})$
$\mathcal{W}_2(\mathbf{p}_x, \mathbf{p}_z) < \mathcal{W}_2(\mathbf{p}_x, \mathbf{p}_{\hat{z}})$
$\mathcal{W}_2(\mathbf{p}_x, \mathbf{p}_z) = \left(\underset{u_{ij}}{\text{minimize}} \sum_{i,j} u_{ij}(x_i - z_j)^2\right)^{1/2}$
$\mathbf{p}_x$
$\mathbf{p}_z$
$\mathbf{p}_{\hat{z}}$
x_i
z_j
$\hat{z}_j$
(b)
p_{z_1}
p_{x_1}
p_{x_2}
p_{x_3}
p_{z_2}
x_1
x_2
x_3
z_1
z_2
(c)
u_{12}
u_{11}
u_{21}
u_{22}
u_{31}
u_{32}
p_{z_1}
p_{x_1}
p_{x_2}
p_{x_3}
p_{z_2}
x_1
x_2
x_3
z_1
z_2

FIGURE 1 (a) The optimal mass transport problem between source $\mathbf{p}_x$ and two target probability histograms $\mathbf{p}_z$ and $\mathbf{p}_{\hat{z}}$ 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. (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 ($\mathcal{W}_2$) increases monotonically as the shift between the central positions of the probability histograms increases [Colour figure can be viewed at wileyonlinelibrary.com]

the bias naturally. More specifically, for $p = 2$, it can be shown that $\mathcal{W}_2^2(\mathbf{p}_x, \mathbf{p}_z) = \mathcal{W}_2^2(\tilde{\mathbf{p}}_x, \tilde{\mathbf{p}}_z) + \|\boldsymbol{\mu}_x - \boldsymbol{\mu}_z\|_2^2$, where $\tilde{\mathbf{p}}_x$ and $\tilde{\mathbf{p}}_z$ are the centred zero-mean probability masses and $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_z$ are the mean values.

3 | REGULARIZATION OF VDA THROUGH THE WASSERSTEIN METRIC

Regularization techniques have been used to reduce the 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 article 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. Consideration of the 2-Wasserstein distance also ensures we keep 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 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 $\mathbf{p}_x = (p_{x_1}, \ldots, p_{x_{km}})^{\mathrm{T}} \in \mathbb{R}^{k^m}$ and $\mathbf{p}_{x_\mathrm{r}} = (p_{x_{\mathrm{r}1}}, \ldots, p_{x_{\mathrm{r}km}})^{\mathrm{T}} \in \mathbb{R}^{k^m}$ represent the vertically concatenated probability vectors of the analysis state and the a priori reference data supported on the "known" matrix $\mathbf{X} = [\mathbf{x}_1 | \ldots | \mathbf{x}_{k^m}] \in \mathbb{R}^{m \times k^m}$, where k represents the number of domain discretizations of the state variable in each dimension and $\mathbf{x}_i \in \mathbb{R}^m$ 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

$$\begin{aligned} \mathcal{J}_{\text{WM-VDA}}(\mathbf{x}, \mathbf{p}_x) = \|\mathbf{x} - \mathbf{x}_\mathrm{b}\|_{\mathbf{B}^{-1}}^2 + \|\mathbf{y} - \mathbf{H}\mathbf{x}\|_{\mathbf{R}^{-1}}^2 \\ + \lambda\, \mathcal{J}_{\mathrm{W2}}\left(\mathbf{p}_x,\ \mathbf{p}_{x_\mathrm{r}}\right), \end{aligned} \tag{9}$$

where $\mathcal{J}_{\mathrm{W2}}(\mathbf{p}_x,\ \mathbf{p}_{x_\mathrm{r}})$ represents the transportation cost associated with the square of the 2-Wasserstein distance between the two probability histograms and λ is a nonnegative regularization parameter, which balances a trade-off between the Euclidean and the Wasserstein cost. There is no closed-form solution to determine this hyperparameter; therefore, it should be estimated empirically through cross-validation experiments.

The analysis state $\mathbf{x}$ at the initial time is an expected value that can be represented as $\mathbf{x} = \mathbf{X}\,\mathbf{p}_x$. Thus, Equation 9 can be expanded as follows:

$$\mathcal{J}_{\text{WM-VDA}}(\mathbf{p}_x) = \|\mathbf{X}\ \mathbf{p}_x - \mathbf{x}_\text{b}\|^2_{\mathbf{B}^{-1}} + \|\mathbf{y} - \mathbf{H}\ \mathbf{X}\ \mathbf{p}_x\|^2_{\mathbf{R}^{-1}} + \lambda\ \mathcal{J}_{\text{W2}}\left(\mathbf{p}_x,\ \mathbf{p}_{x_\text{r}}\right). \tag{10}$$

From the second mass constraint in Equation 8, we have $\mathbf{p}_x = \mathbf{U}\mathbb{1}_{k^m}$ and, setting $\mathcal{J}_{\text{W2}}\left(\mathbf{p}_x,\ \mathbf{p}_{x_\text{r}}\right) = \text{tr}(\mathbf{C}^\text{T}\mathbf{U})$, the above cost function can be expressed in terms of the transportation cost matrix $\mathbf{U}$:

$$\mathcal{J}_{\text{WM-VDA}}(\mathbf{U}) = \|\mathbf{X}\ \mathbf{U}\mathbb{1}_{k^m} - \mathbf{x}_\text{b}\|^2_{\mathbf{B}^{-1}} + \|\mathbf{y} - \mathbf{H}\ \mathbf{X}\ \mathbf{U}\mathbb{1}_{k^m}\|^2_{\mathbf{R}^{-1}} + \lambda\ \text{tr}(\mathbf{C}^\text{T}\mathbf{U}). \tag{11}$$

Let us assume that $\tilde{\mathbf{c}} \in \mathbb{R}^{k^{2m}}$ and $\tilde{\mathbf{u}} \in \mathbb{R}^{k^{2m}}$ denote lexicographic representations of $\mathbf{C} \in \mathbb{R}^{k^m \times k^m}$ and $\mathbf{U} \in \mathbb{R}^{k^m \times k^m}$, respectively, which leads to $\text{tr}(\mathbf{C}^\text{T}\mathbf{U}) = \tilde{\mathbf{c}}^T\tilde{\mathbf{u}}$. 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 the transportation plan in $\mathbf{p}_x = \mathbf{U}\mathbb{1}_{k^m}$ and $\mathbf{p}_{x_\text{r}} = \mathbf{U}^\text{T}\mathbb{1}_{k^m}$ such that $\mathbf{p}_x = \boldsymbol{\Omega}\tilde{\mathbf{u}}$ and $\mathbf{p}_{x_\text{r}} = \boldsymbol{\Lambda}\tilde{\mathbf{u}}$. Here, $\boldsymbol{\Omega} = [\mathbf{I}_{k^m}|\ \mathbf{I}_{k^m}|\ \ldots\ |\mathbf{I}_{k^m}] \in \mathbb{R}^{k^m \times k^{2m}}$ is the horizontal concatenation of the k^m identity matrices and $\boldsymbol{\Lambda} = [\mathbf{e}_1|\ \ldots\ |\mathbf{e}_1|\ \ldots\ |\mathbf{e}_{k^m}|\ \ldots\ |\mathbf{e}_{k^m}] \in \mathbb{R}^{k^m \times k^{2m}}$ is the horizontal concatenation of k^m-dimensional canonical basis, for example, $\mathbf{e}_1 = (1,\ \ldots\ ,0)^\text{T}$ and $\mathbf{e}_{k^m} = (0, \ldots, 1)^\text{T}$.

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

$$\mathcal{J}_{\text{WM-VDA}}(\tilde{\mathbf{u}}) = \frac{1}{2}\tilde{\mathbf{u}}^\text{T}(2\,\boldsymbol{\Omega}^\text{T}\mathbf{X}^\text{T}(\mathbf{B}^{-1} + \mathbf{H}^\text{T}\mathbf{R}^{-1}\mathbf{H})\mathbf{X}\boldsymbol{\Omega})\tilde{\mathbf{u}} + (\lambda\,\tilde{\mathbf{c}}^\text{T} - 2(\mathbf{x}_\text{b}^\text{T}\mathbf{B}^{-1} + \mathbf{y}^\text{T}\mathbf{R}^{-1}\mathbf{H})\mathbf{X}\boldsymbol{\Omega})\tilde{\mathbf{u}}. \tag{12}$$

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

$$\begin{aligned} \tilde{\mathbf{u}}_\text{a} &= \underset{\tilde{\mathbf{u}}}{\text{argmin}}\ \mathcal{J}_{\text{WM-VDA}}(\tilde{\mathbf{u}}), \\ &\text{subject to } \tilde{u}_i \geq 0, \\ &\boldsymbol{\Lambda}\tilde{\mathbf{u}} = \mathbf{p}_{x_\text{r}}, \end{aligned} \tag{13}$$

which can be solved efficiently through interior-point optimization techniques (Altman and Gondzio, 1999). Finally, the analysis state can be obtained as $\mathbf{x}_\text{a} = \mathbf{X}\boldsymbol{\Omega}\tilde{\mathbf{u}}_\text{a}$. It will be noted that, since the analysis state obtained does not satisfy the constraint of the Wasserstein metric exactly, 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 errors, which enables us to characterize the effectiveness of the proposed methodology in comparison with the classic 3D-Var approach. To examine the performance of WM-VDA initially, we focus on two dynamic systems with different levels of sensitivity to their initial conditions, including a first-order linear system and the chaotic Lorenz-63 system. First-order dynamical systems have been the cornerstone in developing the Kalman filter (Kalman, 1960) and have been widely used as a test bed to examine the performance of new filtering techniques (Hazan *et al.*, 2017; 2018). On the other hand, Lorenz-63 has been the subject of numerous experiments to test the performance of new DA techniques under chaotic dynamics (Anderson and Anderson, 1999; Miller *et al.*, 1994; Harlim and Hunt, 2007; Van Leeuwen, 2010; Reich, 2012; Goodliff *et al.*, 2015).

4.1 | First-order linear dynamics

4.1.1 | State-space characterization

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

$$\begin{aligned} \mathbf{x}_{i+1} &= \mathbf{M}\ \mathbf{x}_i + \boldsymbol{w}_i, \\ \mathbf{y}_i &= \mathbf{H}\ \mathbf{x}_i + \boldsymbol{v}_i, \end{aligned} \tag{14}$$

where $\mathbf{M} \in \mathbb{R}^{m \times m}$ 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 $\mathbf{M}$. 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 nonzero mean Gaussian distributions.

4.1.2 | 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 $\mathbf{x}_0 = 10$ and $\mathbf{M} = 0.97$. 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 $\boldsymbol{\omega}_i \sim \mathcal{N}(0.5,\ 1.5)$. The observations are obtained by corrupting the truth at assimilation interval

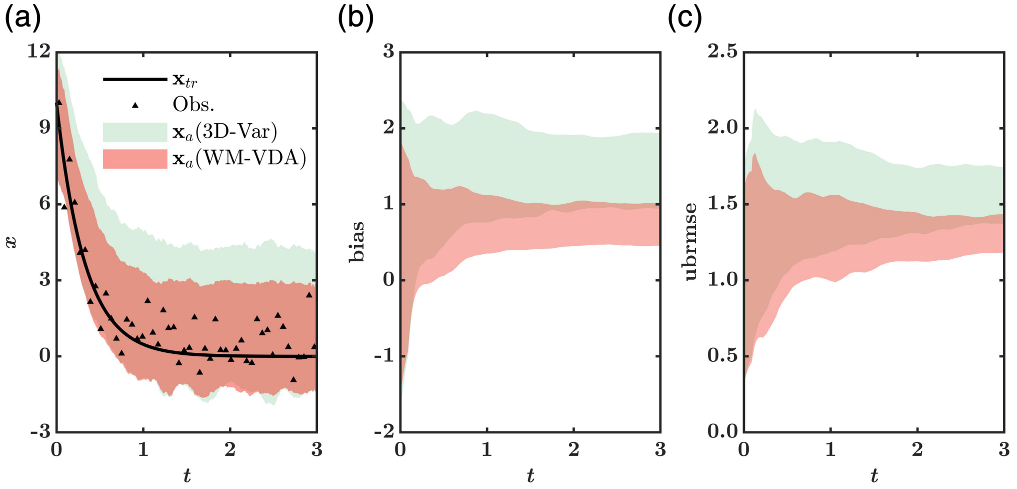

FIGURE 2 Time evolution of the uncertainty range (shaded region) for 3D-Var and WM-VDA, representing (a) the 2.5th and 97.5th percentiles of 50 ensemble members of the analysis state xa and the 95% confidence bound for its (b) bias and (c) ubrmse. The true state (xtr) and observations (Obs.) for an independent simulation are shown in the left panel as well [Colour figure can be viewed at wileyonlinelibrary.com]

$T_{\rm a} = 3\Delta t$ using $\mathbf{v}_i \sim \mathcal{N}(0.25, 0.75)$. To represent the relatively unbiased but highly uncertain a priori reference probability histogram $\mathbf{p}_{x_{\rm r}}$, 500 samples are drawn at each assimilation interval from a Gaussian distribution, where its mean is located at the ground truth and its variance is set to $\boldsymbol{\sigma}^2_{x_{\rm r}} = 3\sigma^2_{\rm b}$, where $\sigma^2_{\rm b} = 1.5$ is the background-error covariance. To solve WM-VDA, the regularization parameter is set to $\lambda = 5$. As will be explained later, this value minimizes the analysis mean-squared error (MSE) empirically. In order to have a robust conclusion about comparison of the proposed WM-VDA scheme with 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 2.5th, and the 97.5th 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,c) improved noticeably for the WM-VDA scheme compared with 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 compared with 3D-Var.

The sensitivity of the quality metrics is also tested for different ranges of assimilation intervals $T_{\rm a} = \{2\Delta t, 5\Delta t, 10\Delta t, 20\Delta t\}$ for both DA schemes (Figure 3a,b). It is found that WM-VDA improves the error quality metrics compared with 3D-Var across the range of assimilation intervals chosen. 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 compared with 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 increase monotonically, however at different rates. In fact, the bias grows faster in 3D-Var than WM-VDA, while the already small gap between the ubrmse values from the two methods shrinks slightly as the assimilation interval grows. Thus, the analysis state bias in WM-VDA seems to be more robust to increased assimilation intervals than 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 λ plays a significant role in the WM-VDA algorithm by making a trade-off between the weighted Euclidean cost and the transportation cost. Recall that larger values of λ push or overfit the analysis state towards the a priori reference probability histogram by neglecting the information content of the 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 λ 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 $\lambda = \{0.1, 5, 50, 1,000\}$ at the first assimilation cycle, for the experimental setting shown 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

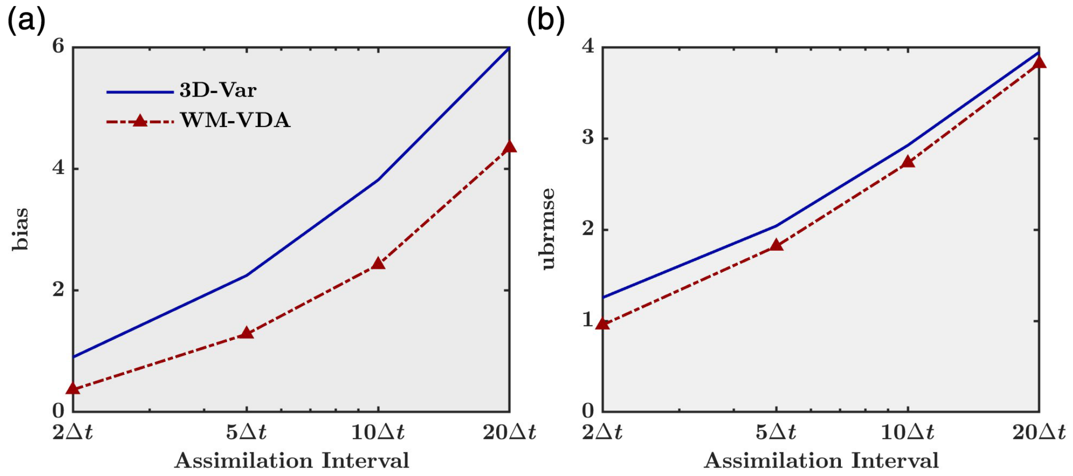

FIGURE 3 Variation of (a) bias and (b) ubrmse for 3D-Var and WM-VDA as a function of assimilation interval [Colour figure can be viewed at wileyonlinelibrary.com]

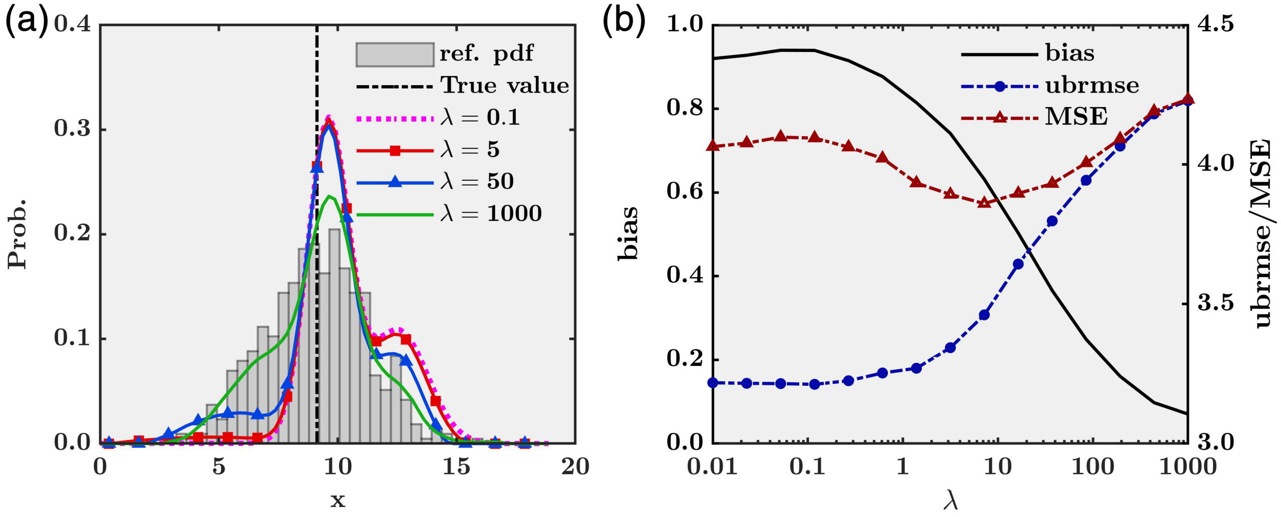

FIGURE 4 (a) Transition of the analysis state probability histogram at the first assimilation cycle and (b) variation of bias, ubrmse, and mean squared error (MSE) of the analysis error as a function of λ, averaged over 50 independent ensemble simulations [Colour figure can be viewed at wileyonlinelibrary.com]

a bimodal distribution as the results approach those of the 3D-Var scheme. However, as the value of λ is increased, the bimodality begins to fade and a more proper geometry of the analysis probability histogram is recovered. At larger values of $\lambda = 1{,}000$, the analysis probability histogram matches perfectly with the a priori reference probability histogram, as the transportation cost dominates the quadratic cost of 3D-Var.

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

4.2 | Lorenz-63

4.2.1 | State-space characterization

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

$$\frac{dx}{dt} = -\sigma(x-y),$$
$$\frac{dy}{dt} = \rho x - y - xz,$$
$$\frac{dz}{dt} = xy - \beta z, \tag{15}$$

where σ, ρ, and β are the Prandtl number, normalized Rayleigh number, and 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 a maximum Lyapunov exponent value of 0.9. For this selection of parameters, there exist three equilibrium points: the origin, which represents the conductive state of no motion, and 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 patterns of convective rolls with different directions of rotation.

4.2.2 | 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 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 with which to construct the reference histogram may not be available at each assimilation cycle. However, a 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 histograms of gauge measurements of surface soil temperature, moisture, or radiosonde data of atmospheric states. To minimize the computational cost of the interior-point optimization algorithm, we solved the problem of determining the 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, 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 every at $10\Delta t$ by corrupting the truth over a simulation period of $T = 0$–20 [t] using a Gaussian error $v_i \sim \mathcal{N}(\boldsymbol{\beta}_y \mathbb{1}_3,\ \sigma_y^2\,\mathbf{I}_3)$, where $\boldsymbol{\beta}_y = 0.15$ and $\sigma_y^2 = 2$. The model is propagated up to $T = 20$ [t] by adding Gaussian noise $\boldsymbol{\omega}_i \sim \mathcal{N}(\boldsymbol{\beta}_{\mathrm{m}} \mathbb{1}_3,\ \sigma_{\mathrm{b}}^2\,\mathbf{I}_3)$ to the model state at every Δt, where $\boldsymbol{\beta}_{\mathrm{m}} = 0.25$ and $\sigma_{\mathrm{b}}^2 = \sqrt{5}$. For setup I, at the assimilation interval of $T_{\mathrm{a}} = 10\Delta t$, 500 samples are drawn from a Gaussian distribution with mean equal to the ground truth and covariance $\sqrt{3}\,\sigma_{\mathrm{b}}^2\mathbf{I}_3$ to construct the unbiased a priori reference probability histogram p_{x_r}. To solve WM-VDA, the regularization parameter is set to $\lambda = 3$ for each dimension by trial and error. However, in setup II, p_{x_r} is constructed by adding a zero-mean Gaussian noise with covariance $\sqrt{3}\,\sigma_{\mathrm{b}}^2\mathbf{I}_3$ to the ground truth over a period $T = 0$–50 [t]. In this case, p_{x_r} has a larger spread than in 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 draw a robust conclusion about comparison of the proposed WM-VDA scheme with 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 2.5th and 97.5th percentiles of the ensemble members from 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 that for 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 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 2.5th and 97.5th percentiles, of the bias and ubrmse is colour-coded in the phase space of the Lorenz system in Figure 6. As is evident, the uncertainty around the computed error statistics gradually shrinks as the DA progresses and passes the duration of the spin-up time. It is seen that the width of the uncertainty in the quality metrics is noticeably narrower in WM-VDA than in 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 $T_{\mathrm{a}} = \{2\Delta t, 5\Delta t, 10\Delta t, 20\Delta t\}$. The expected values of the quality metrics obtained are shown in Figure 7. As expected, for both schemes, bias and ubrmse increase as the assimilation interval grows and, as shown, 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

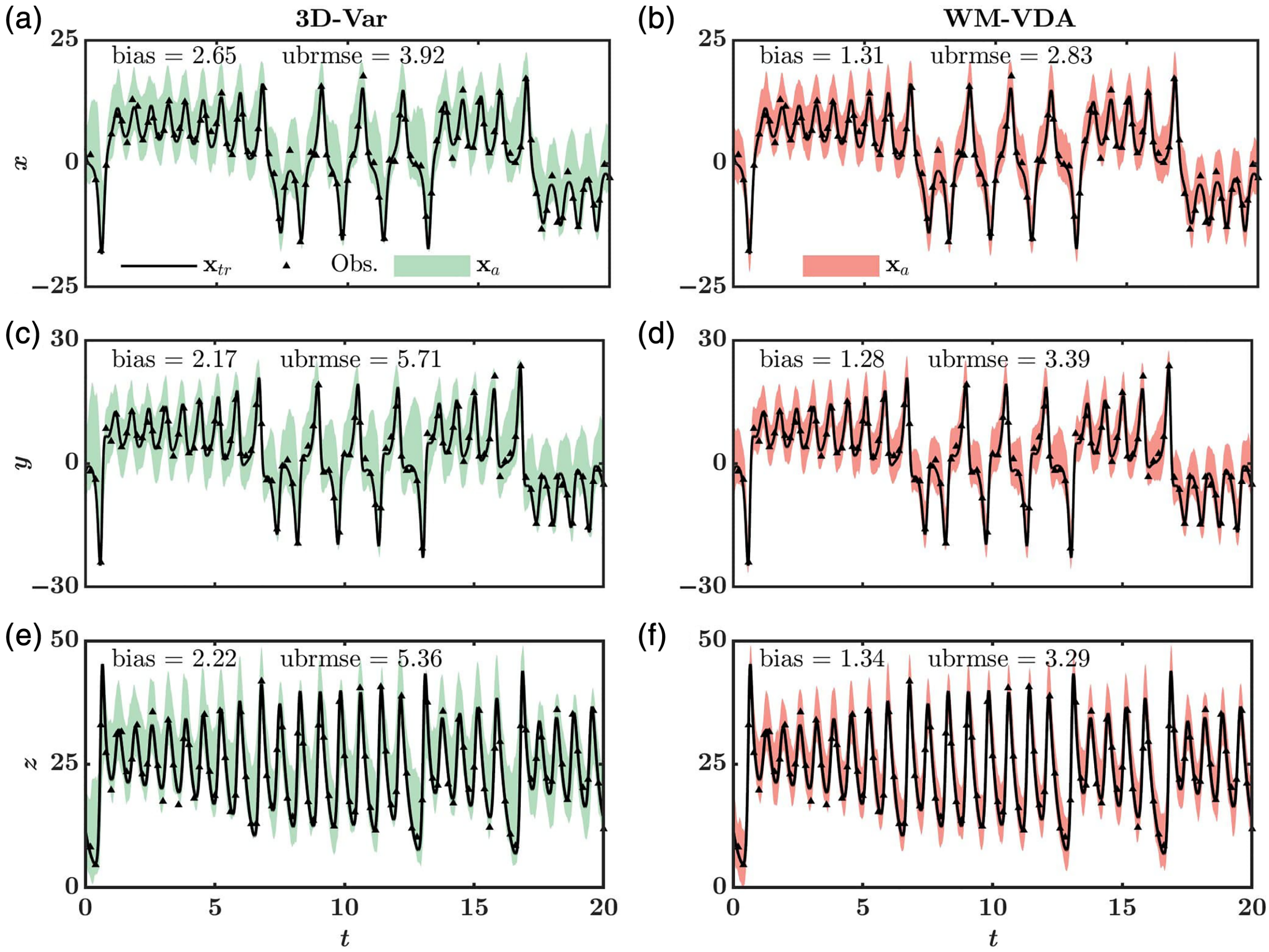

FIGURE 5 Time evolution of the true state (solid lines) of the Lorenz-63 analysis states, observations (Obs.) for one ensemble member and 2.5th and 97.5th percentiles (shaded areas) of the ensemble members for (a, c, e) 3D-Var and (b, d, f) WM-VDA. The results are obtained for 50 independent ensemble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

the ubrmse keeps increasing. This might have stemmed from a very high value of the Lyapunov exponent (0.9) for Lorenz-63. At such a high value, an infinitesimally close trajectories of the state can also deviate significantly as 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 exhibits reduced ubrmse compared with 3D-Var.

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

We also compared the results of setup-II of the WM-VDA scheme with the CDF-matching technique implemented on the Lorenz-63 system with identical assumptions about error structures to those 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

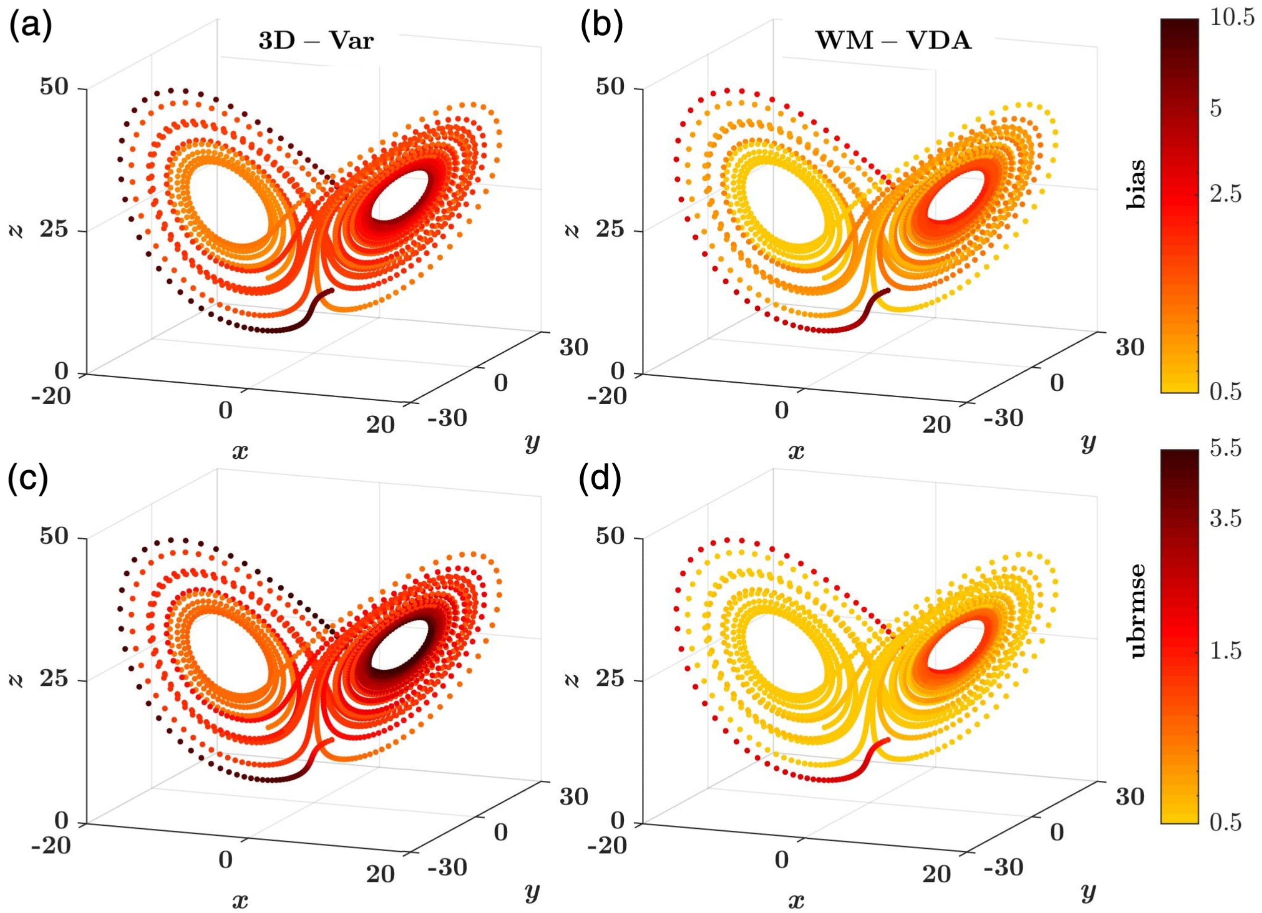

FIGURE 6 Time evolution of the 95% confidence bound of (a,b) bias and (c,d) ubrmse in the phase space of the Lorenz system for 3D-Var and WM-VDA. The uncertainty range is obtained from 50 independent ensemble members for the first experimental setup [Colour figure can be viewed at wileyonlinelibrary.com]

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

	Bias			**ubrmse**		
Methods	***x***	***y***	***z***	***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)

CDF-matching technique by mapping both observations and model state on to the a priori reference dataset using their respective cumulative histograms. To proceed with CDF matching, a cumulative histogram of the reference dataset is constructed by integrating the model in time from an initial state of $(x, y, z) = (3, -3, 12)$ with a time step of $\Delta t = 0.01$ [t] over a period $T = 0$–50 [t] and adding zero-mean Gaussian noise with covariance $\sqrt{3}\,\sigma_{\mathrm{b}}^2 \mathbf{I}_3$, where $\sigma_{\mathrm{b}}^2 = \sqrt{5}$. 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 $v_i \sim \mathcal{N}(\boldsymbol{\beta}_y \mathbb{1}_3,\ \sigma_y^2\ \mathbf{I}_3)$, where $\boldsymbol{\beta}_y = 0.15$ 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 $\boldsymbol{\omega}_i \sim \mathcal{N}(\boldsymbol{\beta}_{\mathrm{m}} \mathbb{1}_3,\ \sigma_{\mathrm{b}}^2\ \mathbf{I}_3)$ to the model state every Δt, where $\boldsymbol{\beta}_{\mathrm{m}} = 0.25$ and $\sigma_{\mathrm{b}}^2 = \sqrt{5}$. At the end of $T = 50$ [t], the model state perturbed with zero-mean Gaussian noise of covariance $\sqrt{3}\,\sigma_{\mathrm{b}}^2 \mathbf{I}_3$ is considered as the initial condition and DA experimentation for the CDF-matching technique is then applied to the next 2000 time steps. At every assimilation time, the observation and model were first mapped on to the reference

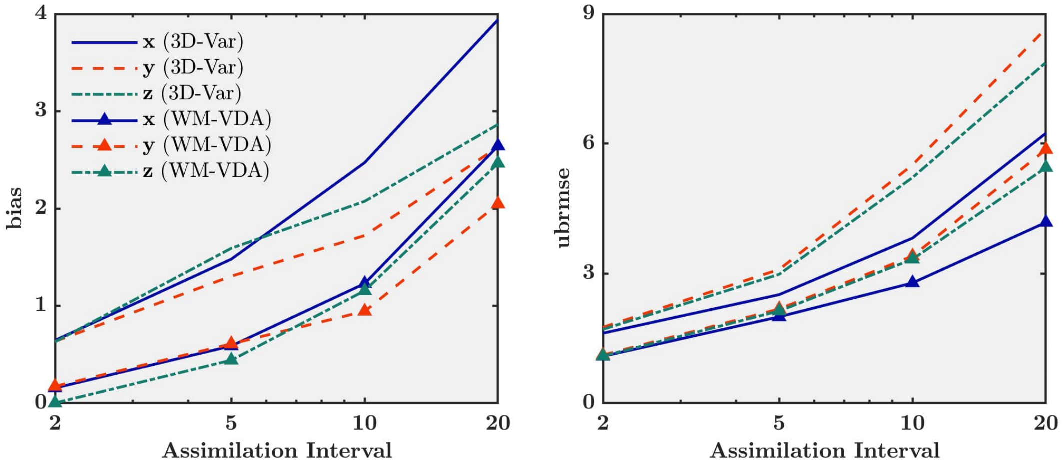

FIGURE 7 Changes of (a) bias and (b) ubrmse as a function of assimilation interval for 3D-Var and WM-VDA in experimental setup I, where (x, y, z) represents the three dimensions of the Lorenz system [Colour figure can be viewed at wileyonlinelibrary.com]

dataset using piecewise linear CDF matching to remove bias, and then the 3D-Var method was utilized to obtain the analysis state using the bias-removed model state and observation. We note that, in comparison with 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 with 28% for WM-VDA, whereas ubrmse increases by 8.6% compared with an overall decrease in ubrmse of 3.3% for WM-VDA. The difference exists primarily due to the way CDF matching maps the biased source of information completely on to the relatively unbiased one. Here, since the reference dataset has a higher uncertainty in terms of variance, CDF matching improves bias compared with 3D-Var at the expense of an increase in ubrmse. Therefore, if the a priori information has uncertainty, as can be expected in a real condition, WM-VDA can be more effective than a statistical CDF-matching technique in bias correction and also provides fertile ground to conduct DA experiments in the probability domain with reduced uncertainty.

5 | SUMMARY AND CONCLUDING REMARKS

In this study, we discussed the concept of Wasserstein metric (WM) regularization of variational data assimilation (VDA) techniques, referred to as WM-VDA. In particular, the classic VDA problem is equipped with a transportation cost that penalizes the mismatch between the probability histograms of the analysis state and relatively unbiased a priori reference data, which can be obtained from in situ observations over a spatial or temporal window. The WM-VDA approach presented does not need any a priori knowledge of the sign of the bias or information about its origin from either model and/or observations, and retrieves it automatically from the a priori reference data. We examined the application of WM-VDA for bias removal in 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 probability histograms (Kolouri *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 characterize the advantages and disadvantages of the proposed approach fully.

As is well understood, in the absence of bias, classic VDA leads to the lowest possible analysis mean-squared error and meets the known Cramér–Rao lower bound (CRLB: Rao *et al.*, 1973; Cramér, 1999). However, in the presence of systematic biases, such a lower bound cannot be met. Even though we demonstrated empirically that, under biased assimilation scenarios, WM-VDA shows improved performance compared with 3D-VAR, future theoretical studies are needed to characterize a closed-form expression for such improvement. We note that, since WM-VDA does not need to attribute the bias specifically proportional to either model or observations, it cannot identify the origin of bias. If such information is needed, future research might be devoted to relating the amount of probability mass transported 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 $\mathcal{O}(k^3 \log k)$, where k is the number of elements in the support set. This hampers applications of 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 reduce its computational cost significantly for high-dimensional problems and could 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 provide new ways for improving bias-aware satellite VDA systems. Of particular interest are satellite soil moisture and precipitation DA (Lin *et al.*, 2015; 2017a), as 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 the 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 concrete ground for solving 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 updated trivially using the likelihood function.

ACKNOWLEDGEMENTS

The first, second, and fourth authors acknowledge a grant from the National Aeronautics and Space Administration (NASA) Terrestrial Hydrology Program (THP, 80NSSC18K1528) through Dr. J. Entin and the New (Early Career) Investigator Program (NIP, 80NSSC18K0742) through Dr A. Leidner and Dr T. Lee. The third and fourth authors also acknowledge support from the National Science Foundation (NSF, 1830418).

ORCID

Sagar K. Tamang https://orcid.org/0000-0001-8301-3576

REFERENCES

Altman, A. and Gondzio, J. (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 Anderson, S.L. (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., Scott, K.A. and Clausi, D.A. (2019) Data fusion and data assimilation of ice thickness observations using a regularisation framework. *Tellus A: Dynamic Meteorology and Oceanography*, 71(1), 1–20.

Asch, M., Bocquet, M. and Nodet, M. (2016) *Data Assimilation: Methods, Algorithms, and Applications*. Vol. 11. Philadelphia, PA: SIAM.

Auligné, T., McNally, A. and Dee, D. (2007) Adaptive bias correction for satellite data in a numerical weather prediction system. *Quarterly Journal of the Royal Meteorological Society*, 133(624), 631–642.

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L. and Peyré, G. (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*, Boston, MA: Springer, pp. 470–482.

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

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

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

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

Courtier, P., Thépaut, J.-N. and Hollingsworth, A. (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, NJ: Princeton University Press.

Crow, W.T., Koster, R.D., Reichle, R.H. and Sharif, H.O. (2005) Relevance of time-varying and time-invariant retrieval error sources on the utility of spaceborne soil moisture products. *Geophysical Research Letters*, 32(24), L24405: 1–5.

Daley, R. (1993) *Atmospheric Data Analysis*, Vol. 2. Cambridge, UK: Cambridge University Press.

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

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

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

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

Dee, D.P. and Uppala, S. (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., Madsen, H. and Rosbjerg, D. (2006) Bias aware Kalman filters: comparison and improvements. *Advances in Water Resources*, 29(5), 707–718.

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

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

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

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

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

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

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

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

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

Hazan, E., Lee, H., Singh, K., Zhang, C. and Zhang, Y. (2018) Spectral filtering for general linear dynamical systems, in. *Advances in Neural Information Processing Systems*, 4634–4643.

Jin, J., Lin, H.X., Segers, A., Xie, Y. and Heemink, A. (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, UK: Cambridge University Press.

Kalnay, E., Li, H., Miyoshi, T., Yang, S.-C. and Ballabrera-Poy, J. (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. *Doklady Akademii Nauk SSSR (NS)*, 37, 199–201.

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

Kolouri, S., Rohde, G.K. and Hoffmann, H. (2018). Sliced Wasserstein distance for learning Gaussian mixture models. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3427–3436. Salt Lake City, UT: IEEE/CV.

Kumar, S.V., Reichle, R.H., Koster, R.D., Crow, W.T. and Peters-Lidard, C.D. (2009) Role of subsurface physics in the assimilation of surface soil moisture observations. *Journal of Hydrometeorology*, 10(6), 1534–1547.

Kumar, S.V., Reichle, R.H., Harrison, K.W., Peters-Lidard, C.D., Yatheendradas, S. and Santanello, J.A. (2012) A comparison of methods for a priori bias correction in soil moisture data assimilation. *Water Resources Research*, 48(3), W03515:1-16.

Kutta, W. (1901) Beitrag zur naherungsweisen integration totaler differentialgleichungen. *Zeitschrift für Angewandte Mathematik und Physik*, 46, 435–453.

Law, K.J. and Stuart, A.M. (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., Wang, Q. and Zhang, L. (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. Los Alamitos, California: IEEE.

Lin, L.-F., Ebtehaj, A.M., Bras, R.L., Flores, A.N. and Wang, J. (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., Ebtehaj, A.M., Flores, A.N., Bastola, S. and Bras, R.L. (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., Ebtehaj, A.M., Wang, J. and Bras, R.L. (2017b) Soil moisture background-error covariance and data assimilation in a coupled land–atmosphere model. *Water Resources Research*, 53(2), 1309–1335.

Liu, Q., Reichle, R.H., Bindlish, R., Cosh, M.H., Crow, W.T., de Jeu, R., De Lannoy, G.J., Huffman, G.J. and Jackson, T.J. (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), 750–765.

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

Lorenc, A.C., Bowler, N.E., Clayton, A.M., Pring, S.R. and Fairbairn, D. (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., Ghil, M. and Gauthiez, F. (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*, 666–704.

Motamed, M. and Appelo, D. (2019) Wasserstein metric-driven Bayesian inversion with applications to signal processing. *International Journal for Uncertainty Quantification*, 9(4), 395–414.

Ning, L., Carli, F.P., Ebtehaj, A.M., Foufoula-Georgiou, E. and Georgiou, T.T. (2014) Coping with model error in variational data assimilation using optimal mass transport. *Water Resources Research*, 50(7), 5817–5830.

Pauwels, V.R., De Lannoy, G.J., Hendricks Franssen, H.-J. and Vereecken, H. (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. and Cuturi, M. (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., Houser, P.R., da Silva, A. and Bosilovich, M.G. (2001). Results from global land-surface data assimilation methods. In: *AGU Spring Meeting Abstracts*. Boston, Massachusetts: American Geophysical Union (AGU).

Rao, C.R. (1973) *Linear Statistical Inference and its Applications*, Vol. 2. New York, NY: Wiley.

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

Reich, S. and Cotter, C. (2015) *Probabilistic Forecasting and Bayesian Data Assimilation*. Cambridge, UK: Cambridge University Press.

Reichle, R.H. and Koster, R.D. (2004) Bias reduction in short records of satellite soil moisture. *Geophysical Research Letters*, 31(19), L19501:1-4.

Reichle, R.H., Koster, R.D., Dong, J. and Berg, A.A. (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., Koster, R.D., Liu, P., Mahanama, S.P., Njoku, E.G. and Owe, M. (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), D09108:1-14.

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

Rubner, Y., Tomasi, C. and Guibas, L.J. (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*, Vol. 55, pp. 58–63. New York, NY: Birkäuser.

Snyder, C., Bengtsson, T., Bickel, P. and Anderson, J. (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., Künsch, H.R., Nerger, L., Potthast, R. and Reich, S. (2019) Particle filters for high-dimensional geoscience applications: a review. *Quarterly Journal of the Royal Meteorological Society*, **145**(723), 2335-2365.

Villani, C. (2003) *Topics in Optimal Transportation*, Vol. 58. Providence, Rhode Island: American Mathematical Society.

Villani, C. (2008) *Optimal Transport: Old and New*, Vol. 338. Heidelberg, Germany: Springer Science & Business Media.

Villarini, G., Mandapaka, P.V., Krajewski, W.F. and Moore, R.J. (2008) Rainfall and sampling uncertainties: a rain gauge perspective. *Journal of Geophysical Research: Atmospheres*, 113(D11), D11102:1-12.

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

Wang, X., Barker, D.M., Snyder, C. and Hamill, T.M. (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., Derber, J., Collard, A., Dee, D., Treadon, R., Gayno, G. and Jung, J.A. (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.

How to cite this article: Tamang SK, Ebtehaj A, Zou D, Lerman G. Regularized variational data assimilation for bias treatment using the Wasserstein metric. *Q J R Meteorol Soc*. 2020;146:2332–2346. https://doi.org/10.1002/qj.3794