Quantifying the Representativeness Errors Caused by Scale Transformation of Remote Sensing Data in Stochastic Ensemble Data Assimilation

Feng Liu, Zebin Zhao, and Xin Li, *Senior Member, IEEE*

***Abstract*—Representativeness error caused by scale transformation (REST) is an intrinsic property of data assimilation, as assimilating new observations likely involves the fusion of multisource and multiscale data. Earlier studies focused on specific cases and failed to obtain a general concept. This study attempts to achieve a further understanding of REST in both theory and practice. Based on scale-related definitions and formulations, the statistical RESTs of observation errors and analysis errors are deduced in stochastic ensemble data assimilation. Experiments based on ensemble Kalman filter are conducted to validate the interpretability of the proposed formulations. A synthetic experiment uses the stochastic Lorenz model as the forecasting operator, and a real-world experiment employs a simple biosphere model as the forecasting operator and uses a series of mixed ground-based and remote sensing soil moisture observations. The results confirm that REST should be proportional to the scale difference when assimilating direct observations and both system states and observations are homogeneous processes. Due to the nonlinearity in modeling, assimilation, and scale transformation, increasing RESTs are found if the scale of the observation is much larger than that of the state space, or multiscale observations are added into the assimilation system. Quantifying REST improves the understanding of uncertainty in data assimilation, but further studies on REST are required in both theory and practice, for example, REST correlates with other errors in forcing, parameters, and models, and introduces an observation operator to assimilate indirect observations.**

***Index Terms*—Heihe watershed allied telemetry experimental research (HiWATER), scaling, lorenz model, remote sensing, SiB2 model, soil moisture, stochastic process, uncertainty, wireless sensor networks.**

Manuscript received September 12, 2021; revised November 24, 2021, January 7, 2022, and February 1, 2022; accepted February 4, 2022. Date of publication February 9, 2022; date of current version March 2, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 42130113 and Grant 41801270, in part by the National Science and Technology Major Project of China's High-Resolution Earth Observation System under Grant 21-Y20B01-9001-19/22, in part by the Basic Research Innovative Groups of Gansu Province, China under Grant 21JR7RA068, and in part by the Foundation for Excellent Youth Scholars of "Northwest Institute of Eco-Environment and Resources," CAS. *(Corresponding authors: Feng Liu; Xin Li.)*

Feng Liu and Zebin Zhao are with the Key Laboratory of Remote Sensing of Gansu Province, Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000, China (e-mail: liufeng@lzb.ac.cn; zhaozebin@lzb.ac.cn).

Xin Li is with the National Tibetan Plateau Data Center, State Key Laboratory of Tibetan Plateau Earth System, Resources and Environment, Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing 100101, China, and also with the CAS Center for Excellence in Tibetan Plateau Earth Sciences, Chinese Academy of Sciences, Beijing 100101, China (e-mail: xinli@itpcas.ac.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3149957

I. Introduction

REPRESENTATIVENESS errors in earth observations, modeling, and assimilations mainly refer to errors caused by inconsistencies in spatial-temporal resolution among different geophysical observation and models [1]–[4], differences in observation techniques and retrieval methods used for the same geophysical variables [5], and deficiencies in available models compared to ideal models [3]. The first type of representativeness error has also resulted from spatial scale (for brevity, "scale" hereafter refers to the spatial scale) transformation among multisource observations. Stemming from spatial heterogeneities and irregularities [6], [7] in geographical variables across different scales, this error is related either to significant dynamic process variations in land surface systems, such as saturated hydraulic conductivity [8]–[10], soil structures [11], long-term water balance simulations [12], the spatial correlations of surface soil moisture [14], or to the physics of remote sensing, including the radiative transfer process of vegetation [13] and the validity of Planck's law [15].

Representativeness error caused by scale transformation (REST) has received increasing attention in the fields of data assimilation and remote sensing. Data assimilation forms a unified and generalized framework by combining earth system modeling and observations [16], and consequently, it is necessary to address changes in scales among different model units and observation supports in data assimilation; thus, data assimilation is an ideal tool for characterizing REST. Traditionally, scale and scale transformation are implicit in data assimilation systems. However, studies can also be conducted by dividing the system states into resolved and unresolved portions [2], using restriction and prolongation operators [17], considering the resolution-based relationships between the system state and observation [3], deriving a new gain matrix for the system state and observation in true and inaccurate forecasting spaces [4], and assimilating cloud-profiling radar observations [18]. Experimental studies investigating REST in remote sensing, such as multiscale validations of soil moisture observations [19], [20], solar radiation measurements [21], [22], and upscaling carbon flux measurements [23], are continuously conducted. These explorations are constructive; however, they cannot lead to a unified understanding of REST, which first requires theoretical studies that explicitly consider scale transformation.

Stochastic data assimilation mainly refers to data assimilation employing stochastic process based geophysical variables. The advantage of a stochastic framework is that it produces an infinite evolution probability distribution space for both system state [24] and observation. Stochastic data assimilation is an ongoing problem [25]–[27]. Recently, a stochastic data assimilation framework containing the explicit expression of scale transformation was formulated [28]. This new framework provides a promising approach to addressing REST by defining scale and introducing scale into the posterior probability distribution function (PDF) of data assimilation.

This study attempts to formulate the statistical REST in stochastic ensemble data assimilation and further conducts synthetic and real-world experiments by assimilating ground-based and remote sensing observations to trace REST. To conduct more practical and generalized experiments, we attempt to develop approaches that simulate any heterogeneity in geophysical variables at diverse scales; therefore, cases in which observations vary in isotropy or anisotropy are the main focus. In Section II, basic knowledge regarding how to characterize REST is first introduced. Rigorous definitions of scale and scale transformation are presented. The corresponding likelihood and REST in observation error and analysis error are formulated in a stochastic data assimilation system. The methods used to simulate REST, including the adopted forecasting models, data types, sampling, and development environment of data assimilation systems, are presented in Section III. Sections IV and V present two experiments evaluating REST when assimilating multiple scales observations. The detailed analysis of the results, and the corresponding discussion and conclusions are provided in Sections VI and VII, respectively.

II. REST in Stochastic Ensemble Data Assimilation

To quantify REST, specifying the scale should be given top priority, and the scale-related errors need to be mathematically identified. Note that all formulas hereinafter are deduced in the sense of scale, and the partial derivatives are with respect to the elements of scale s or state vector $X(s)$. Therefore, the formulas do not conflict with the existing expressions because we only consider scale transformation. For more details about the derivations in this section, please refer to [28].

A. Scale-Related Definitions

To improve the typical expression defining scale in terms of distance, a two-dimensional definition of scale was proposed [28] as follows:

$$s = \iint_A dx_1 dx_2 \tag{1}$$

where scale s is a Lebesgue measure [29], [30] with respect to the observation footprint or model unit A in the two-dimensional real number space R^2. This definition is related to the ability to refer to scales in multidimensional spaces and distinguish more complicated changes in scales.

If two different scales exist, i.e., s_1 and s_2, scale transformation can be defined based on the Lebesgue integration by substitution as, (6) shown at bottom of next page.

$$s_2 = \iint_{A_2} dy_1 dy_2 = \iint_{A_1} |J(x_1, x_2)|\, dx_1 dx_2 \xrightarrow{J=\mathrm{diag}(\xi,\xi),\ \xi\in R}$$

$$s_2 = |J| \iint_{A_1} dx_1 dx_2 = \xi^2 s_1 \tag{2}$$

where $J(x_1, x_2)$ or J is the Jacobian matrix, and A_1 and A_2 are the arbitrary observation footprints or model units of s_1 and s_2, respectively. Therefore, Jacobian matrix $J(x_1, x_2)$ represents the geometric transformation from A_1 to A_2. The right part of (2) indicates that the transformation between s_1 and s_2 is one-dimensional, i.e., s_1 and s_2 are similar geometries.

Accordingly, if the scale transformation is one-dimensional, the Ito process based geophysical variables can be expressed as follows:

$$dV = \varphi(s)\, ds + \sigma(s)\, dW(s) \tag{3}$$

where s, $W(s)$, $\varphi(s)$, and $\sigma(s)$ represent the scale, Brownian motion, scale-based drift and variance, respectively.

Stochastic data assimilation is formulated by introducing the following stochastic process based state vector:

$$dX = \varphi_X(s)\, ds + \sigma_X(s)\, dW(s) \tag{4}$$

and/or observation

$$dY = \varphi_Y(s)\, ds + \sigma_Y(s)\, dW(s). \tag{5}$$

In related studies on stochastic data assimilation [24], [27], stochastic processes (4) and (5) used other physical elements, such as time t, as independent variables.

B. Likelihood in Stochastic Data Assimilation

Assuming that the nonlinear observation operator $H(s, x)$ is scale dependent with the continuous partial derivatives H_s, H_x, and H_{xx}, based on the Ito lemma [31], we have $X \sim N(X_0 + \int_{s_0}^{s_X} \varphi_X(s)ds, \int_{s_0}^{s_X} \sigma_X^2(s)ds)$, $Y \sim N(Y_0 + \int_{s_0}^{s_Y} \varphi_Y(s)ds, \int_{s_0}^{s_Y} \sigma_Y^2(s)ds)$, and $H(s_Y, X(s_Y)) = H(s_0, X_0) + \int_{s_0}^{s_Y} [H_s + H_x\varphi_X + \frac{1}{2}H_{xx}\sigma_X^2]du + \int_{s_0}^{s_Y} H_x\sigma_X dW(u)$. Then, the likelihood in general terms can be deduced in Eq. (6).

$$p(Y|X) = N\left(Y(s_Y) - \left[H(s_X, X(s_X)) + \int_{s_X}^{s_Y}\left(H_s + H_x\varphi_X + \frac{1}{2}H_{xx}\sigma_X^2\right)du\right], \int_{s_X}^{s_Y} H_x^2\sigma_X^2 du\right) \tag{6}$$

In particular, if the observation has the same physical quantity as the state vector (assimilating direct observation), i.e., $H(s,x)=x$, $H_s=0$, $H_x=1$, and $H_{xx}=0$, the likelihood is deduced as follows, (6) shown at bottom of next page.

$$p(Y|X) = N\left(Y(s_Y) - \left[X(s_X) + \int_{s_X}^{s_Y} \varphi_X du\right], \int_{s_X}^{s_Y} \sigma_X^2 du\right). \tag{7}$$

Regardless of the heterogeneity, namely, $\varphi_X=0$ and $\sigma_X=1$, when transforming the scale-dependent state vector from the state space to observation space, (7) can be reduced as follows:

$$p(Y|X) = N\left\{Y(s_Y) - X(s_X), |s_Y - s_X|\right\}. \tag{8}$$

Let the conditions used to build (8) be *Condition Sim*. Condition Sim refers to the simplest situation in which the observation is direct, and scale transformation only involves the Gaussian process, viz., $H=E$, $\varphi_X=0$, $\sigma_X=1$.

Compared with ordinary data assimilation based on deterministic models, stochastic data assimilation provides a more applicable and general framework for studying REST in integrated and time-continuous synthetic experiments.

C. REST in Analysis Error of the Ensemble Kalman Filter (EnKF)

As a widely accepted assimilation algorithm, the EnKF provides an operational way to evaluate the effect of REST on analysis error. The analysis step of EnKF is given by

$$\begin{cases} K = P^f H^T \left(H P^f H^T + R\right)^{-1} \\ X^a = X^f + K\left(Y - H\left(X^f\right)\right) \\ P^a = (I - KH) P^f = \frac{1}{N-1}\left(X^a - \overline{X^a}\right)\left(X^a - \overline{X^a}\right)^T \end{cases} \tag{9}$$

where K, P^f, R, P^a, X^a, and X^f are Kalman gain matrix, forecasting error matrix, observation error covariance matrix, analysis error matrix, analysis state vector, and forecasting state vector, respectively.

Assuming that P^f is not involved in scale transformation and that R is constant for simplicity (also a common choice for many data assimilation developments), then K is free of scale transformation and can be regarded as constant in the analysis step. X^a is also a stochastic process provided by $X^a = X^f + K\Sigma$, where $\Sigma = Y - H(X^f) \sim p(Y|X) = N\{Y(s_Y) - X(s_X), |s_Y - s_X|\}$. Considering the Condition Sim, $X^a \sim N(X^f + K(Y(s_Y) - X(s_X)), K|s_Y - s_X|K^T)$ and

$$P^a = K|s_X - s_Y|K^T. \tag{10}$$

Equation (10) indicates that, under the condition that Y is direct observation of X and both X and Y are homogeneous processes, analysis error is proportional to the degree of scale transformation.

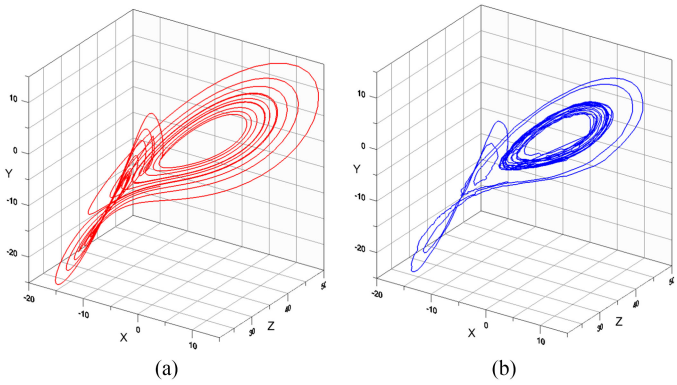

Fig. 1. Three-dimensional curves of the (a) Lorenz and (b) SLMs, where $\sigma=10$, $\rho=28$, and $\beta=8/3$, with the stochastic parameters $g_1=g_2=g_3=1/2$.

III. Materials

A. Common Land Data Assimilation Framework

All experiments are conducted with a common data assimilation software package ComDA [33]. ComDA is used to fuse multisource earth observations in various land surface applications. ComDA is capable of performing nonlinear and non-Gaussian data assimilation methods (including a variety of nonlinear Kalman filters and particle filters), general multiple observation operators (such as radiative transfer models) and multiple model operators (Common Land Model, SiB2, etc.). The corresponding software is developed with parallel and distributed computing technologies, and implemented in Linux-based programming languages.

B. Stochastic Lorenz Model (SLM)

An SLM [see Fig. 1(b)] is a stochastically perturbed form of Lorenz63 [see Fig. 1(a)] equations [25] as follows:

$$\begin{cases} dX_1 = \sigma(X_2 - X_1)\,dt + g_1 dW_1 \\ dX_2 = (\rho X_1 - X_2 - X_1 X_3)\,dt + g_2 dW_2 \\ dX_3 = (X_1 X_2 - \beta X_3)\,dt + g_3 dW_3 \end{cases} \tag{11}$$

where σ, ρ, and β are the Prandtl numbers, normalized Rayleigh number, and nondimensional wavenumber, respectively. State vector $X=\{X_i,\ i=1,2,3\}$ is formed by the stochastic process, where $W_i,\ i=1,2,3$, are the independent Brownian motions, and $g_i,\ i=1,2,3$, are the variance coefficients.

The SLM performs well in regular ensemble data assimilation strategies [25]. In this study, SLM is employed as a forecasting operator and is assumed to be free of scale transformation.

C. SiB2

SiB2 [34] is the second version of the simple biosphere model [35]. SiB is a typical biophysical process and empirical hybrid model used to simulate the transfer processes of energy, water, carbon, mass, and momentum among the atmosphere and biosphere, in which physical state variables including radiation fluxes, carbon cycle, and water stores are mainly considered. In SiB2, the satellite data are further considered, and the carbon

cycle, water exchanges, and land surface reflectance simulations are improved by introducing more realistic models. SiB2 has been widely applied in land surface modeling to understand energy and carbon flux [36], exchanges [37], and hydrology [38]. SiB2 was also applied as a biophysical module in a regional data assimilation system [39].

In SiB2, soil moisture is assumed to be measured in three divided layers (surface, root zone, and depth zone) corresponding to water evaporation, root uptake, and recharge. The corresponding governing equations are mainly based on Richards' equation, which works well in the finer running spatial units but has difficulty capturing large regional hydrological processes. Meanwhile, Richards' equation only formulates the gravity-driven vertical flow through the unsaturated zone and fails to incorporate horizontal flow through the saturated zone. These characteristics suggest that soil hydrological processes described in SiB2 are sensitive to the local geology and land surface heterogeneities and, consequently, vary with scales.

D. Data Types

1) Synthetic Noise: A spatially correlated Gaussian random field is generated by the geoR package (http://leg.ufpr.br/geoR/), in which the spatial mean, nugget, and sill of the geostatistical semivariogram model are 0, 1, and 1, respectively.

The following multiscale airborne–ground-based observations are used.

2) SoilNET: Ground-based soil moisture measurements are provided, with each node configured with four sensors at the depths of 4, 10, 20, and 40 cm, an observation interval of 10 min and a spatial support of 10 cm.

3) COsmic-Ray Soil Moisture Observing System (COSMOS) [43]: To measure the neutrons generated by cosmic rays within air, soil, and other materials, a COSMOS probe was installed to estimate the time series of area-average soil moisture retrieval [46]. The corresponding observation interval and spatial support are 60 min and a footprint circle of 600 m, and the effective measured depth is from 76 cm (dry soils) to 12 cm (saturated soils).

4) Polarimetric L*-Band Multibeam Radiometer (PLMR) Flight Multiangle Observations [44]:* Through airborne passive microwave radiation of the land surface with dual-polarization and multiangled observation. PLMR generates soil moisture retrieval measured at a depth of 5 cm and a higher spatial resolution of 700 m compared with coarser remote sensing products.

E. Metrics

1) Analysis Errors and Average Analysis Errors (AAEs): According to (10), analysis error in each analysis step reflects the change of scale in EnKF-based data assimilation, which will be conducted in the following experiments. The AAE is further provided to examine how REST affects analysis error during the entire assimilation, i.e.,

$$\overline{o^i} = \frac{1}{n}\sum_{k=1}^{n} o_k^i \tag{12}$$

where $\overline{o^i}$ and o_k^i denote the AAE and analysis error at the *k*th analysis step of the *i*th element in the state vector, respectively, and n denotes the total analysis time. Here, we assume that each pair of elements in the state vector are independent. Using many samples, AAE provides a simple statistic for surveying the REST in ensemble data assimilation and checks the degree of scale change in (10).

2) Differences Between Data Probability Spaces: Since both the states and observations obey given probabilities, quantifying the distances among these probability spaces can contribute to evaluating the impact of REST on the evolutions of states and observations.

Both the correlation coefficient and Kullback–Leibler (KL) divergence [40] are introduced to distinguish the data probability spaces with different scales. The correlation coefficient is used as a consistency check between different assimilation estimation vectors. The KL divergence is used to define the distance between two different probability spaces. If F_1 and F_2 represent two discrete probability distributions and their probability density functions are f_1 and f_2, respectively, then the KL divergence from F_1 to F_2 is $D_{\mathrm{KL}}(F_1, F_2) = \sum_{i=1}^{n} f_1(i)\log[f_1(i)/f_2(i)]$. Since the KL divergence is asymmetric, i.e., $D_{\mathrm{KL}}(F_1, F_2) \neq D_{\mathrm{KL}}(F_2, F_1)$, the distance between two data probability spaces F_1 and F_2 can be written as

$$\begin{aligned} d_{\mathrm{KL}}(F_1, F_2) &= \frac{1}{2}\left[D_{\mathrm{KL}}(F_1, F_2) + D_{\mathrm{KL}}(F_2, F_1)\right] \\ &= \frac{1}{2}\sum_{i=1}^{n}\left[f_1(i) - f_2(i)\right]\log\left[f_1(i)/f_2(i)\right]. \end{aligned} \tag{13}$$

This distance follows the prime properties of a general distance: 1) nonnegativity means that $d_{\mathrm{KL}}(F_1, F_2) \geq 0$; 2) symmetry means that $d_{\mathrm{KL}}(F_1, F_2) = d_{\mathrm{KL}}(F_2, F_1)$; and 3) definiteness means that if $d_{\mathrm{KL}}(F_1, F_2) = 0$, then $F_1 = F_2$.

IV. Synthetic Experiment

A. Design of the Experiment

By using (4) and (5) to represent the system state and observation in the analysis step, respectively, REST can be characterized in a data assimilation system. Equation (6) presents a general expression of REST in which the diverse heterogeneities across different scales can be formulated by defining the scale-based drifts φ and variances σ.

The simplest isotropic case follows the Condition Sim, where the geophysical variables change evenly among different scales (8), i.e., only Gaussian random errors are involved in scale transformations between the state space and observation space. Additionally, the anisotropic case indicates more general conditions under which scale-dependent variations should be considered. Here, we design a special case of heterogeneity to present the nonlinear variations in scale differences due to the nonlinear dynamics of models and assimilation algorithms.

Fig. 2 illustrates how to simulate these two cases in a stochastic data assimilation system with EnKF. For simplicity, we assume that only scale transformation is involved in the analysis step

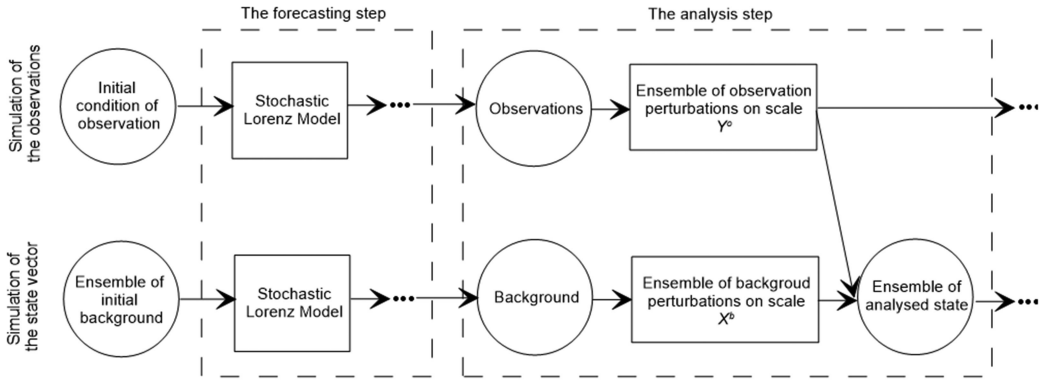

Fig. 2. Flowchart of simulating RESTs in a stochastic data assimilation system with the EnKF.

and that the observations are "direct" measurements of the state vector, i.e., there is no observation operator in this data assimilation system. SLM is used to produce both the observations and state vector with different initial values. These two components can be regarded as geophysical variables [(4) and (5)], and their ensembles with respect to different scales are generated in the analysis step.

In the case of isotropic change, the scale-dependent geophysical variables are defined as $dV = dW(s)$, and random Gaussian distributed error occurs only when the geophysical variables transform from one scale to another. Perturbing the observations and state vector is performed by simply setting $Y^{o,p} \sim N(Y^o, s)$ and $X^{b,p} \sim N(X^b, s)$, where $Y^{o,p}$ and $X^{b,p}$ are the perturbed observations and state vector, respectively. Additionally, in the anisotropic case, we employ the definition of $dV = e^{-1.5+s}ds + dW(s)$ as the scale-dependent geophysical variables to investigate the variations; therefore, mapping the state vector from the state space to the observation space should consider the sampling space $X^{b,p} \sim N(X^b + e^{-1.5+s}, s)$. We add the synthetic noise into this sampling space to ensure the introduction of spatially correlation with respect to scale. It should be noted that this sampling space is only an example and cannot be generalized to the concept of anisotropy. In fact, modeling all anisotropic scenarios in one case is an impossible mission.

B. Results

There are 500 running steps in total in the isotropic experiment, and an observation is added into the data assimilation system every 50 steps. The number of ensemble members of the background field in the forecasting step and the number of perturbations in the analysis step are both 100. The experiment shown in Fig. 3 assumes that the scales of the state space and two different observation spaces (obs1 and obs2) are 1, 1, and 2, respectively.

As shown in Fig. 3, two different assimilation results (assim1 and assim2) are obtained with respect to the observations in two different observation spaces (obs1 and obs2). The first result comprises the ensemble means (blue curves), forecasting and analysis errors (blue histograms), and observations (blue dots). The same elements shown in red constitute the second result.

TABLE I
AAEs OF EACH ENSEMBLE IN THE ANALYSIS STEPS

		X_1	X_2	X_3
Isotropic	assim1	1.1982	1.4586	1.9939
experiment	assim2	1.4944	2.1063	2.8530
Anisotropic	assim1	1.2774	1.6247	2.2937
experiment	assim2	1.5915	2.4042	3.2562

Each assimilated result is found to simulate the truth accurately, but the second result presents a more apparent trend of deviation from the true state vector (such as periods from step 250 to 300, from step 350 to 400, etc.). This phenomenon is due to scale mismatches between the state space and observation space, i.e., scale in the second observation space is twice the scales in the state space and the first observation space, leading to obvious observation errors in mapping the state vector from the state space to the observation space and to disagreement with the true state vector.

The forecasting or analysis errors of the ensemble can also provide similar information. As clearly shown in Fig. 3, errors increase as the number of simulation steps increases until a new observation is added into the data assimilation system, and then the errors significantly decrease. Meanwhile, errors of the second ensemble are larger than those of the first ensemble at most simulation times, indicating that the greater the discrepancy between the scales of observation and state space in the analysis step, the larger the errors will be at the forecasting step.

More detailed information regarding REST can also be provided by the AAE. The corresponding numbers are listed in Table I. In the first two rows, assim2 clearly produces a larger error regardless of whether the element of the state variable is X_1, X_2, or X_3. This result indicates that if a variable is disturbed by larger scale-dependent unbiased noise, simulating this variable requires sampling the space with larger variance.

AAE is related to (10). In this isotropic experiment, we regard Y as a direct observation of the state vector; therefore, AAE should be proportional to $|s_Y - s_X|$. Fig. 4 shows AAEs of ensembles $X = \int_{s_X}^{s_Y} dW(s)$, with $s_X = 1$ and $s_Y = 2,\ 3,\ 4,\ 5,\ 10,\ 20,\ 30$, and 50. An ideal result is that all AAE curves are parallel to or overlap with the reference line. However, the gradually upward-moving trend of these lines is unstoppable

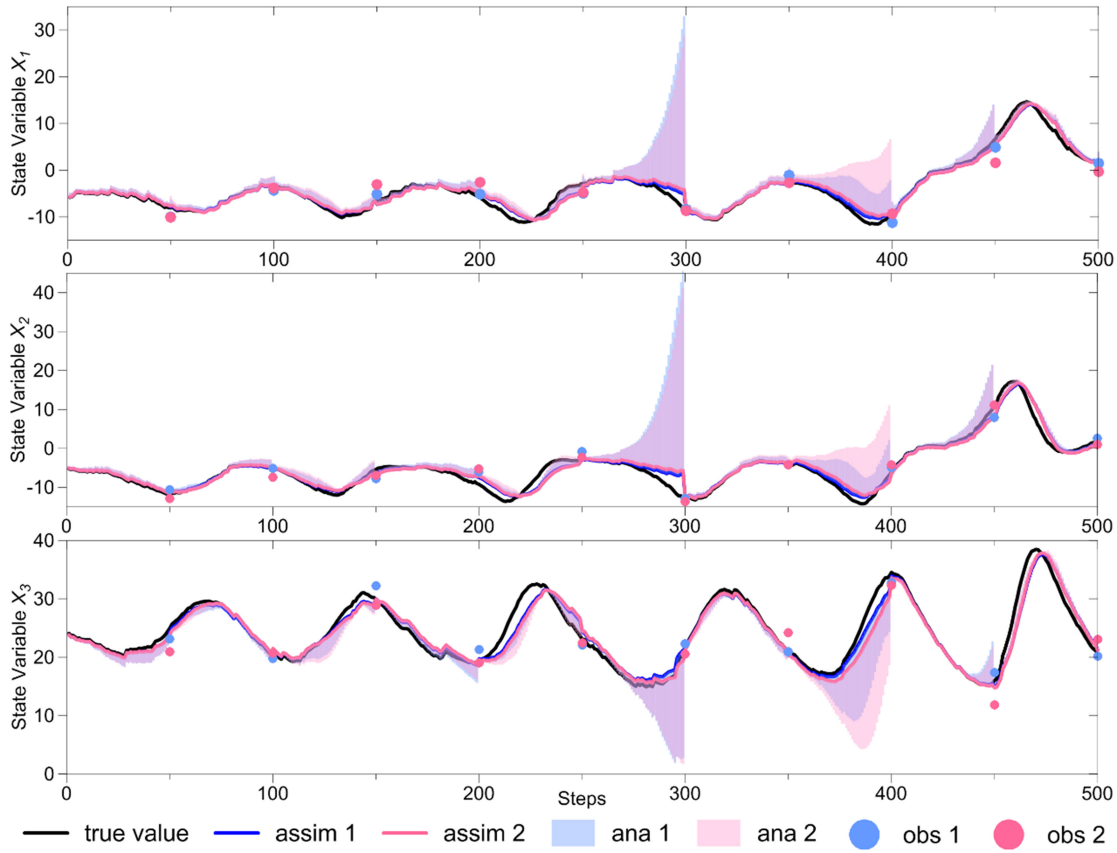

Fig. 3. Isotropic experiment simulating RESTs in a stochastic data assimilation system (assim = assimilation; ana = analysis or forecasting error; obs = observation).

as the scale increases. All AAE curves present a slope similar to that of the reference when the scale is smaller than 10, but the slope is steeper when the scale is larger than 10. This phenomenon contradicts (10). Another set of figures also challenges the prior conclusion that AAE should be proportional to the scale differences between the state and observation space. The average difference between $s_Y = 1$ and $s_Y = 2$, $s_Y = 2$ and $s_Y = 3$, $s_Y = 3$ and $s_Y = 4$, and $s_Y = 4$ and $s_Y = 5$ is 1.24, which is very close to the scale difference $|s_Y - s_X| = 1$. However, this value increases to 16.12 between $s_Y = 5$ and $s_Y = 10$ and further increases to 78.97 between $s_Y = 10$ and $s_Y = 20$ and between $s_Y = 20$ and $s_Y = 30$, both of which are far from their scale differences. We assume that except for the limitation of sampling, this inconsistency between the simulation results when $s_Y > 10$ and (10) is mainly due to the over-perturbation involved in the scale transformation. Thus, if the noise variances are much larger than the values of the geophysical variables, this theory is no longer true. In this experiment, the range of the state variables is approximately –15 to 40; thus, perturbations with a variance greater than 10 may significantly influence the dynamic processes of the state and thereby produce increasing AAEs.

The anisotropic experiment uses the same configuration settings as the isotropic experiment. Fig. 5 shows the assimilated results of the anisotropic experiment. The corresponding simulations greatly differ from the true value (such as the period from step 200 to 300 for all state variables). The deviations are mainly caused by the scale variation of the geophysical variables and are proportional to the intensity of the variation. As shown in Fig. 5, since assim2 has stronger scale variation, its assimilation results are more likely to depart from the true values than those of assim1. Accordingly, the analysis errors of assim2 are larger than those of assim1. Another significant difference is that the analysis errors in this experiment are much larger than those in the isotropic experiment. For example, the analysis errors in step 300 of state variable X_1 are approximately 30 in the isotropic experiment and more than 40 in the anisotropic experiment. Furthermore, based on (7), the scale-dependent dynamic process φ_X can impact the likelihood. If the state vector is more likely to be influenced by scale transformation, a larger the error will be produced. Thus, if a geophysical variable is liable to be scale variant, its assimilated results will be no better than those of scale-invariant variables.

The AAEs of assim1 and assim2 are listed in the third and fourth row of Table I. Assim2 still produces larger variances in this experiment. However, compared with the isotropic experiment, AAEs stem not only from the scale-dependent unbiased perturbation but also from the scale-variant dynamic process of the state vector.

The AAEs of ensembles $X = X(s_X) + e^{-1.5+s}|_{s=s_X}^{s=s_Y} + \int_{s_X}^{s_Y} dW(s)$, with $s_X = 1$ and $s_Y = 2, 3, 4, 5, 6, 7, 8$, and 9 are shown in Fig. 6. All lines representing AAEs have a slope

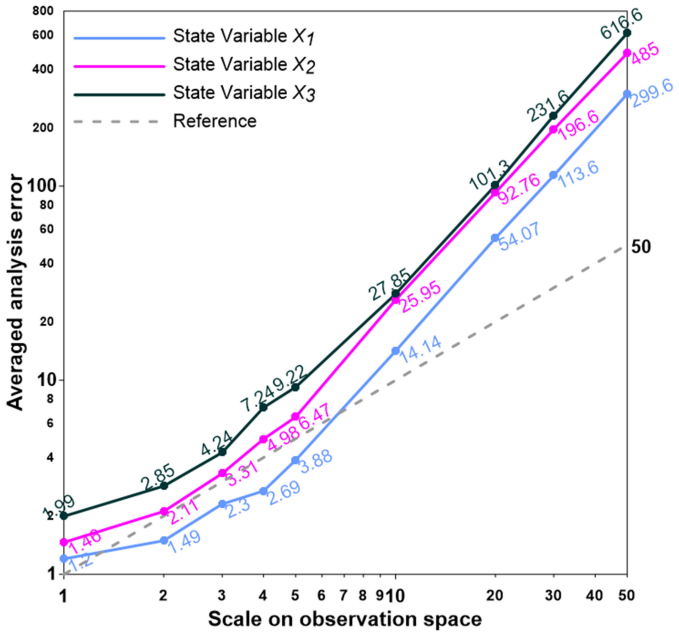

Fig. 4. AAEs of ensembles ($\overline{P^a}$) versus the scale of observation space (s_Y) on logarithmic coordinates. The gray dashed line references the function $s_Y = \overline{P^a}$.

similar to that of the reference when the scale is no more than 5, but the slope is higher when the scale is larger than 5. The upward trends of these lines are steeper than those shown in Fig. 4. We assume that in addition to over-perturbation, this phenomenon is attributed to the nonlinear scale variation of the geophysical variable. When mapping the state vector from the state space to observation space, the value may become very different from its original value, resulting in inaccurate simulations and extraordinary errors.

V. Real-World Experiment

A. Study Area and Data Processing

The study area of the real-world experiment is the $962\ \mathrm{m} \times 962\ \mathrm{m}$ Yingke-Daman irrigation district in the midstream region of the Heihe River Basin, which is the core experimental region of Heihe Watershed Allied Telemetry Experimental Research (HiWATER) [41]. Multiscale airborne–ground-based observations have been conducted in HiWATER including Eco-Hydrological Wireless Sensor Networks (EHWSN) [42], COSMOS, and PLMR observations. There are 50 SoilNET nodes in total installed in EHWSN, a COSMOS probe was installed in the center of the study area, and several synchronized PLMR flights have been conducted. The configurations of the observations are illustrated in Fig. 7.

The observed time span from June 23, 2012 00:00 to August 9, 2012 16:00 (GMT+8, and similar hereinafter) is selected as the experimental period. During this period, there are 1000 COSMOS observation times in total, and soil moisture data from SoilNET are hourly averaged to match COSMOS data in terms of collection frequency. PLMR data are only available from noon on June 30, 2012; July 7, 2012; July 26, 2012; and August 2, 2012. The corresponding datasets are provided by CASEarth Poles.[1] Note that spatial averages may introduce extra uncertainties into the assimilation system; therefore, only one SoilNET node and one pixel of PLMR are selected for direct comparison with COSMOS (see Fig. 7).

TABLE II
Assimilation Strategy for Twin Tests Adding Soil Moisture in Three Layers in SiB2

	Surface	Root zone	Deep zone
Reference test	data i	data ii	data iii
1st control test	data i	data iv	data iii
2nd control test	data vi	data ii	data iii

In the study area, the root and deep zones are 0.48 and 1.5 m, respectively [47]. Therefore, the observations of SoilNET at 4 and 20 cm are used to represent the soil moisture in the first two layers in SiB2, and assimilating the observations at 40 cm into the soil moisture in the third layer is performed for reference only due to the lack of data.

B. Design of the Experiment

Based on the multiscale observations in the study area, the following six time series are created to characterize REST in a real-world land data assimilation system.

1) Data i describes the soil moisture data observed by the hourly averaged SoilNET at a depth of 4 cm.
2) Data ii describes the hourly averaged SoilNET data at a depth of 20 cm.
3) Data iii describes the hourly averaged SoilNET data at a depth of 40 cm.
4) Data iv describes the soil moisture data observed by COSMOS.
5) Data v is a time series of alternating data i and iv, i.e., taking turns to sampling the observations from data i and iv at each time.
6) Data vi is a time series of combining data v and available PLMR data.

Regarding these time series as both forcing and observation data of soil moisture in the three layers in SiB2, twin tests (see Table II) are conducted to examine the evolution of REST when assimilating the observations at different scales (i.e., the support scales are SoilNET with a circle of radius 0.1 m, COSMOS with a circle of radius 600 m, and PLMR with a square of side length 700 m, respectively) and depths. The reference test uses data i–iii; then, the first control test is the same as the reference but replaces data ii with data iv, and the second control test uses data vi to replace data i. Compared with the reference test, the first control test aims to assimilate larger-scale soil moisture at the root zone, the second control test tries to simulate assimilation of multisource and multiscale observations. Note that it is reasonable to employ COSMOS data for both surface and root layers because the vertical support of COSMOS is an effective depth, which certainly covers the measured ranges of SoilNET nodes at the depths of 4 and 20 cm. All tests are examined according

[1][Online]. Available: http://www.tpdc.ac.cn [45]

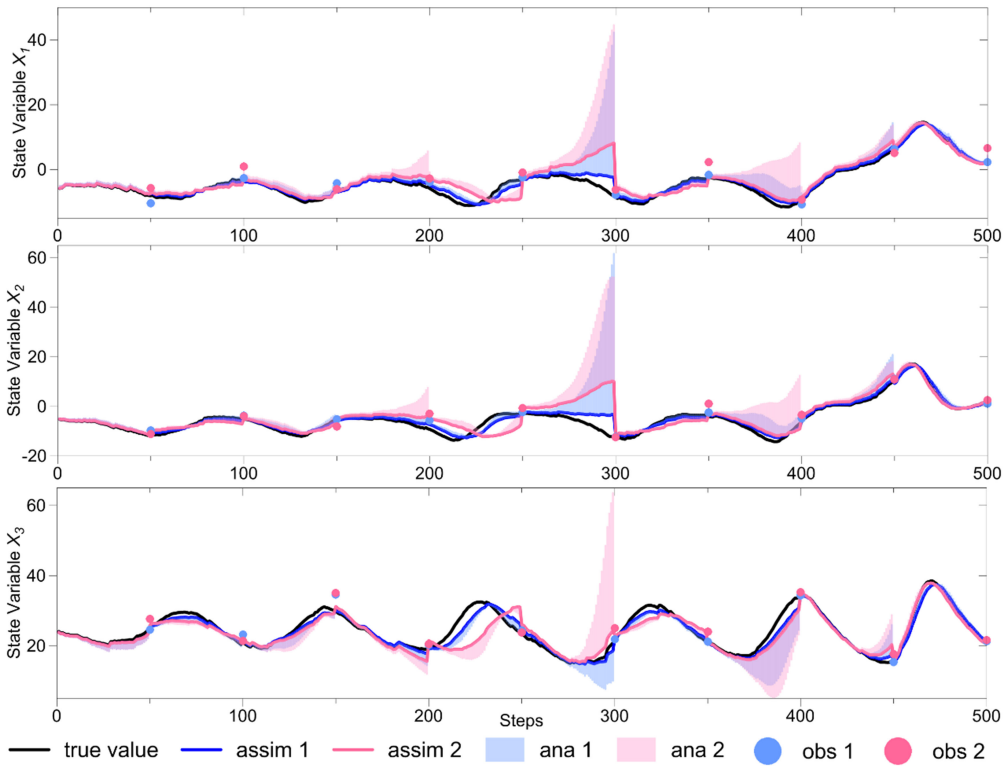

Fig. 5. Anisotropic experiment simulating RESTs in a stochastic data assimilation system (assim = assimilation; ana = analysis or forecasting error).

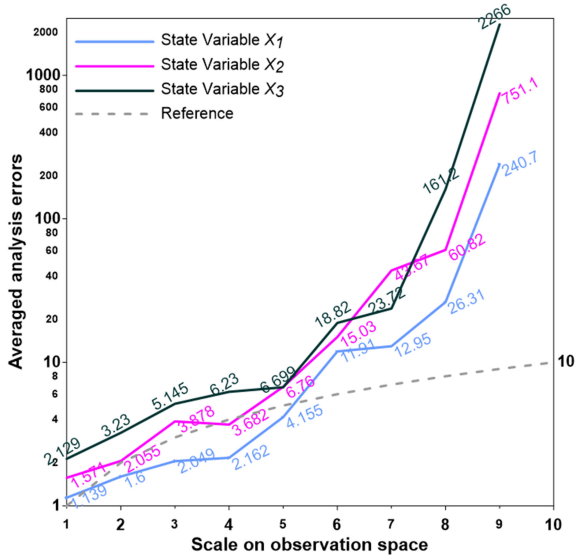

Fig. 6. AAEs of the ensembles ($\overline{P^a}$) versus the scale of the observation space (s_Y) on logarithmic coordinates. The gray dashed line references the function $s_Y = \overline{P^a}$.

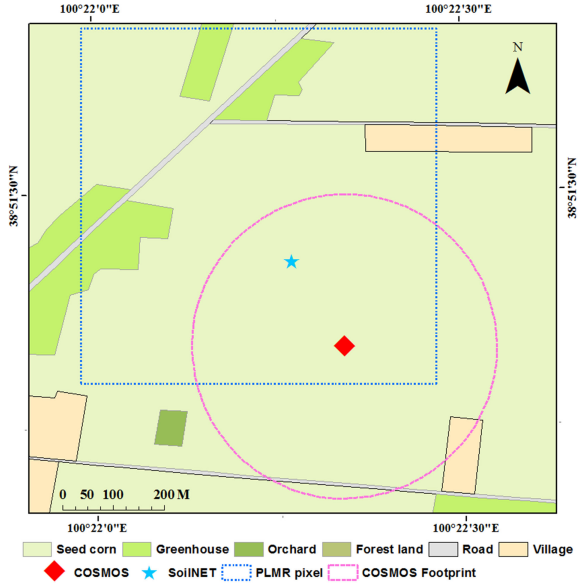

Fig. 7. Configurations of multiscale observations in the study area during an intensive observation period of HiWATER. The support scales of SoilNET, COSMOS, and PLMR are 0.1, 600, and 700 m, respectively.

to the correlation coefficient, KL divergence, and analysis errors produced in terms of REST.

C. Results

There are 1000 observation times in total during the experimental period both for SoilNET and COSMOS. SiB2 is also implemented during this period. The interval of assimilating a new observation is 49, which ensures that the SoilNET and COSMOS data are alternately added into data assimilation system. Observations are also assimilated when PLMR data are available. The numbers of ensemble members and perturbations are both 100.

Fig. 8 presents the corresponding results of the reference and control tests. Clearly, with no available observations, neither of

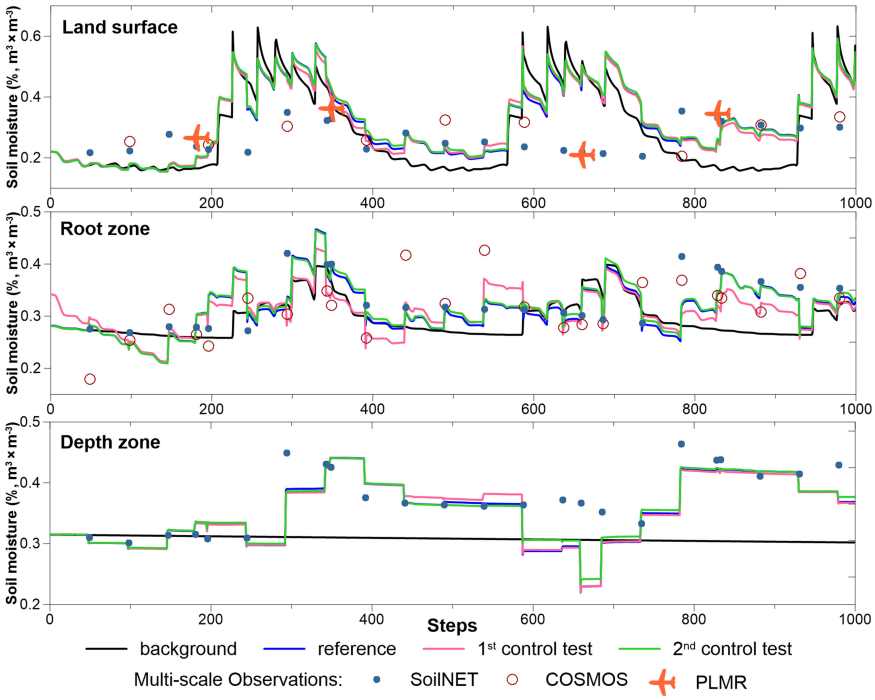

Fig. 8. Real-world experiment in the Heihe River Basin assimilating multiscale observations and evaluating the corresponding RESTs. (Assimilation algorithm: EnKF, forecasting operator: SiB2, running frequency: 1 h, assimilation frequency: 49 h.)

TABLE III
CORRELATION COEFFICIENTS OF THE ASSIMILATION RESULTS BETWEEN THE REFERENCE AND CONTROL TESTS[a]

	Surface			Root zone			Depth zone		
	FH	SH	Total	FH	SH	Total	FH	SH	Total
1st control test	0.997	0.994	0.996	0.911	0.775	0.878	0.996	0.995	0.995
2nd control test	0.999	0.999	0.999	0.998	0.988	0.995	0.999	0.996	0.996

[a]FH = The First Half of the results, SH = The Second Half of the results.

the assimilation results simulates the trend of the background in the depth zone. Therefore, this case is not covered in the following text. However, the soil moisture assimilation results in the surface and root zone present similar trends to that of the reference and have obvious characteristics when adding multiscale observations. In the root zone, the soil moisture forcing data in the reference and the first control test are, respectively, collected by SoilNET and COSMOS, which indicates that the initial values of these two assimilation systems are different. The trends of reference and first control test are consequently distinguished from each other at the beginning of the experimental period, but they may have an increasing resemblance and become continually closer until the next observation is added into system. In addition, irrigation period and meteorological forcing such as rainfall are not considered in this experiment; therefore, in a few steps, the soil moisture values are slightly larger and result in the corresponding overestimates in the land surface layer in SiB2.

The correlation coefficients (see Table III) between the twin tests for the different layers also provide remarkable information. Clearly, each control test strongly agrees with the reference, except the estimates of the first control test at the root zone, which uses COSMOS data as larger-scale observations. Moreover, the correlation coefficient of the second half of this estimate is significantly lower than that of the first half, implying that, respectively, assimilating the observations at different scales can result in increasing deviation between these sequential assimilation estimates. Another important detail in the information concerns the differences between the reference and control tests, which become larger over time for all layers. That is, the correlation coefficients of the second half are lower than those of the first half even though the observations are the same, for example, the reference and first control tests in the land surface layer, or the reference and second control tests in the root zone layer.

TABLE IV
AAES ($M^3 \times M^{-3}$) AND THEIR KL DIVERGENCES OF EACH REAL-WORLD TEST

		Surface	Root zone	Depth zone
Reference test	AAE	2.3688	5.8041	3.9062
1st control test	AAE	2.3754	6.0001	3.9211
	KL	0.0020	0.0059	0.0003
2nd control test	AAE	2.4016	5.8448	3.9205
	KL	0.0222	0.0032	0.0001

The corresponding analysis errors and KL divergences can indicate how REST impacts the assimilation system. In Table IV, it is found that the AAEs in both the land surface

layer of the second control test and the root zone layer of the first control test are larger than their counterparts. It is certain that assimilating observations at different scales can result in increasing REST, followed by larger analysis errors. However, the increase in analysis errors is less remarkable than that in the synthetic experiments and is also inconsistent with (10). The KL divergences between reference and control tests agree with the above results. The probability spaces of the two mentioned assimilation analysis errors are distant from that of the reference, which proves that REST can further influence the probability distribution of the analysis error.

VI. Discussion

A comprehensive understanding of REST requires studies in both theory and applications. Based on the proposed theory of scale and scale transformation, this study puts forward the formulations of REST in observation errors and analysis errors in ensemble assimilation, and further introduces a series of synthetic experiments and real-world experiments to evaluate the REST in data assimilation systems.

A. Synthetic Experiments

The advantages of simulating and quantifying REST using a stochastic process are that the related experiment is controllable, and the results are general. If deterministic operators are introduced, the corresponding results could be influenced by the deterministic processes, which would not lead to a complete understanding of REST under the data assimilation framework.

Two synthetic experiments are conducted to simulate scale-dependent geophysical variables in a stochastic data assimilation system and quantify RESTs by setting the observation spaces with various scales. Basically, the proposed theory of formulating REST in data assimilation is confirmed. However, we further determine that this theory cannot provide a reasonable explanation for why the AAEs became remarkable when the scale of the observation is much larger than that of the state space. Although this problem may be caused by the sampling techniques, unreasonable scale-dependent perturbances, and significant scale variations in the dynamic processes, this limitation indicates that the REST theory needs to be refined in our future work. Furthermore, the interpretability scope of this theory should be further defined.

Both the scale variations of the dynamic process and stochastic disturbance can result in heterogeneity across different scales. Compared with the stochastic perturbances, the impact of the scale variation in the dynamic process on heterogeneity is crucial, which may result in accumulating REST and significant errors. The stochastic state vector in the anisotropic experiment, i.e., $dX = e^{-1.5+s}ds + dW(s)$, may be too simple and incorrect for some true geophysical variables, but this special case performs well in producing assimilated results if the scale variation in the geophysical variable is considered. Defining σ_X based on scale s or Brownian motion $W(s)$ produces stochastic heterogeneity. However, this special case is not included in this study because this heterogeneity is beyond the current understanding of earth observation, modeling, and assimilation, and the synthetic experiment involves complicated solutions for stochastic differential equations, which does not appear to facilitate drawing the final conclusions.

An apparent limitation of this synthetic experiment is that not all anisotropic cases can be simulated. Stochastic data assimilation is a nonlinear system and is simulated by the ensemble method. However, the methods presented in this study can simulate only limited cases that produce REST, i.e., one isotropic case and one anisotropic case in which the geophysical variable monotonically increases with the change in scale. In contrast to the isotropic case, which is based on only one definition, anisotropic cases present infinite variations. Simulating all anisotropic cases in one synthetic experiment is impossible, but we believe that similar conclusions can be obtained regardless of how the anisotropic case is defined.

B. Real-World Experiments

REST in the real-world experiment performs differently from that in the synthetic experiment. The following factors may contribute to understanding this phenomenon.

1) Theory Validation: There are three observations at different scales involved in the twin tests, namely SoilNET (0.1 m), COSMOS (600 m), and PLMR (700 m). However, only SoilNET and COSMOS data can be used to validate the REST in the analysis error since PLMR is not continuously observed. We consequently only consider the estimates of the root zone between the reference and the first control test. In Fig. 8, the estimate of the first control test is considerably different from that of the reference, which is also presented in Tables III and IV, and its corresponding correlation coefficients and KL divergences are inconsistent with those of the second test. In Table IV, compared with the second test, the AAE of this estimate is slightly larger, which proves that (10) is basically correct since the REST in the analysis error increases with an increasing difference between scales. However, the current results cannot indicate that REST is proportional to the scale difference. Due to the lack of available observations at different scales in the real-world settings, we cannot design an intensive test to quantify the relationship between REST and scale difference. More importantly, (10) is deduced under the condition of assimilating direct observations and homogeneous system states, which is largely different from the real world. Therefore, (10) needs to be generalized. Following these numerical experiments, the related mathematical understanding is consequently needed.

2) Nonlinear Performances: Due to the nonlinearity of the data assimilation system, REST should display fast divergence or convergence. However, no remarkable fluctuation is found during the experimental period. There is also no chaotic behavior in the real-world assimilation since the estimate is not sensitive to the initial states (see Fig. 8, estimates of the root zone in the reference and first control tests). These phenomena prove that the data assimilation system is long-term controllable and predictable.

Another nonlinear characteristic of assimilation may be introduced by the related variables, that is, introduction of scale-changed observations to update a specific state can impact its

related states. Considering the land surface layer in the first control test or the root zone in the second control test, although both use the same data as that in the reference, the corresponding estimates gradually deviate from the reference, as shown in both Fig. 8 and Table III (the correlation coefficient of the second half is smaller than that of the first half). The AAEs of the corresponding estimates are further larger than those of the reference (see Table IV). This deviation may also be caused by the random perturbances, but the increasing AAE proves that REST is propagated from one state of assimilating observation at a different scale to the related variables. The relationships between correlated variables, such as the soil moisture at these two layers, have been certainly formulated in physics, but developing an understanding of the nonlinear propagation process of REST between the correlated variables still needs further studies.

3) Assimilating Multiscale Observations: It is widely recognized that introducing multisource observations can, in principle, improve the predictability of an assimilation system. However, multiple observations generally lead to scale issues since they are collected at different resolutions. The surface layer estimate of the second control test is designed to examine the corresponding REST. Table IV summarizes that both the AAE and its KL divergence in this test are larger than those in the first test when assimilating multiscale observations at the surface layer. However, if using the same observations as those in the reference in the other layers, these two figures become lower than those of the first test. We therefore presume that assimilating multiscale observations can result in increasing REST; however, this presumption should be further proven in future theoretical studies.

We only examine the strategy of assimilating one observation in one analysis step, and its formulation and related experiments are further conducted. The simultaneous assimilation of multiple observations is not included in this study since the corresponding theory and experiments are more complicated. For example, the extra correlation between the multiple observations should be considered. We plan to address this problem in future work.

In the real-world experiments, the errors originating from the horizontal distances among the centroids of different observations are not considered. Studying REST implies that all centroids of the observations/simulations are coincident, which usually cannot be met due to the lack of data. This issue may result in additional uncertainties and requires intensive investigation in future studies.

C. Intercomparison

The synthetic experiment can create a series of pseudo-scale-continuous observations to evaluate REST and deduce general conclusions, whereas real-world experiment only uses limited observations but presents more practical significance for studies of REST.

REST has been explicitly formulated under a stochastic ensemble data assimilation framework. However, the interpretability and limitation of this theory still need to be validated by observations and experiments. Considering that an adequate concept of data assimilation that directly matches REST is lacking, two elements, i.e., AAE and KL divergence, were used to represent the REST estimates and the distance between different REST PDFs. However, REST still cannot be fully understood in this study since other errors, for example, random perturbances of ensembles, are involved. Moreover, REST in a real-world experiment may encounter further interference from the uncertainties of forcing, parameters, and dynamic forecasting processes. RESTs are correlated with other errors and cannot be simply identified; therefore, understanding REST in data assimilation still requires in-depth studies.

Based on theory and the synthetic experiment, scale transformation is shown to cause remarkable variation in REST. However, this variation is restricted to a specific region in the real-world experiment, considering the control mechanism of the assimilation system and the weak feedback of the land surface process, and seemingly, REST has a limited impact on the entire system. However, this finding does not preclude the urgency of studying REST. In the real-world experiment, only one scale-change state is involved. In reality, the scale issue is inevitable in both the forcing data and parameters, and models are also not free of scale variation. In comprehensive assimilation studies, scale transformation can be found in each interaction between data and models; therefore, the corresponding RESTs constitute more key components of uncertainty, and the mechanism by which they influence the final results needs further study. The study of REST can effectively improve the understanding of uncertainty in data assimilation.

This study considers only direct observations. Introducing indirect observations into stochastic data assimilation requires the formulation of stochastic observation operators (such as stochastic radiative transfer models), which requires an analysis or numerical solutions of the stochastic operators. However, a stochastic model suitable for the stochastic data assimilation framework that produces the desired outputs is unavailable. We will attempt to address this problem in the future.

VII. Conclusion

Introducing stochastic processes and models in data assimilation does not change its basic formulations. The results confirm that the proposed theory holds under the circumstantial setting of a small change in scale but do not offer sufficient evidence that this theory works well when the scale changes by any significant degree.

This study provides scale-related definitions of REST in both observation error and analysis error, and designs synthetic and real-world assimilation experiments to validate the corresponding formulations. We further prove that introducing multiscale observations into data assimilation can produce considerable representativeness errors, and ignoring REST is unreasonable because scale transformation is an intrinsic property of earth observations, modeling, and assimilations. This study provides a feasible approach to simulating and quantifying REST in data assimilation when introducing multisource observations including remote sensing and ground-based measurements.

REFERENCES

[1] C. Lorenc, "Analysis methods for numerical weather prediction," *Quart. J. Roy. Meteorol. Soc.*, vol. 112, no. 474, pp. 1177–1194, 1986.

[2] T. Janjić and S. E. Cohn, "Treatment of observation error due to unresolved scales in atmospheric data assimilation," *Monthly Weather Rev.*, vol. 134, no. 10, pp. 2900–2915, Oct. 2006.

[3] P. J. van Leeuwen, "Representation errors and retrievals in linear and nonlinear data assimilation," *Quart. J. Roy. Meteorol. Soc.*, vol. 141, no. 690, pp. 1612–1623, 2015.

[4] D. Hodyss and N. Nichols, "The error of representation: Basic understanding," *Tellus A, Dyn. Meteorol. Oceanogr.*, vol. 67, no. 1, 2015, Art. no. 24822.

[5] F. Liu and X. Li, "New spatial upscaling methods for multi-point measurements: From normal to p-normal," *Comput. Geosci.*, vol. 109, pp. 247–257, 2017.

[6] X. Li, "Characterization, controlling, and reduction of uncertainties in the modeling and observation of land-surface systems," *Sci. China Earth Sci.*, vol. 57, no. 1, pp. 80–87, 2014.

[7] T. Zhao *et al.*, "Soil moisture experiment in the Luan River supporting new satellite mission opportunities," *Remote Sens. Environ.*, vol. 240, 2020, Art. no. 111680, doi: 10.1016/j.rse.2020.111680.

[8] D. Giménez, W. J. Rawls, and J. G. Lauren, "Scaling properties of saturated hydraulic conductivity in soil," *Geoderma*, vol. 88, no. 3, pp. 205–220, 1999.

[9] H. Vereecken, R. Kasteel, J. Vanderborght, and T. Harter, "Upscaling hydraulic properties and soil water flow processes in heterogeneous soils: A review," *Vadose Zone J.*, vol. 6, no. 1, pp. 1–28, 2007, doi: 10.2136/vzj2006.0055.

[10] G. A. Narsilio, O. Buzzi, S. Fityus, T. S. Yun, and D. W. Smith, "Upscaling of Navier–Stokes equations in porous media: Theoretical, numerical and experimental approach," *Comput. Geotechnics*, vol. 36, no. 7, pp. 1200–1206, 2009.

[11] H. Lin, H. Flühler, W. Otten, and H.-J. Vogel, "Soil architecture and preferential flow across scales," *J. Hydrol.*, vol. 393, no. 1, pp. 1–2, 2010.

[12] R. Merz, J. Parajka, and G. Blöschl, "Scale effects in conceptual hydrological modeling," *Water Resour. Res.*, vol. 45, no. 9, 2009, Art. no. W09405.

[13] S. Jacquemoud *et al.*, "PROSPECT+SAIL models: A review of use for vegetation characterization," *Remote Sens. Environ.*, vol. 113, pp. S56–S66, 2009.

[14] D. Ryu and J. S. Famiglietti, "Multi-scale spatial correlation and scaling behavior of surface soil moisture," *Geophys. Res. Lett.*, vol. 33, no. 8, 2006, Art. no. L08404.

[15] X. Li, A. Strahler, and M. Friedl, "A conceptual model for effective directional emissivity from nonisothermal surfaces," *IEEE Trans. Geosci. Remote. Sens.*, vol. 37, no. 5, pp. 2508–2517, Sep. 1999.

[16] X. Li, F. Liu, and M. Fang, "Harmonizing models and observations: Data assimilation in Earth system science," *Sci. China Earth Sci.*, vol. 63, no. 8, pp. 1059–1068, 2020, doi: 10.1007/s11430-019-9620-x.

[17] M. Bocquet, L. Wu, and F. Chevallier, "Bayesian design of control space for optimal assimilation of observations. Part I: Consistent multiscale formalism," *Quart. J. Roy. Meteorol. Soc.*, vol. 137, no. 658, pp. 1340–1356, 2011, doi: 10.1002/qj.837.

[18] M. D. Fielding and O. Stiller, "Characterizing the representativity error of cloud profiling observations for data assimilation," *J. Geophys. Res., Atmos.*, vol. 124, no. 7, pp. 4086–4103, 2019.

[19] W. T. Crow *et al.*, "Upscaling sparse ground-based soil moisture observations for the validation of coarse-resolution satellite soil moisture products," *Rev. Geophys.*, vol. 50, no. 2, 2012, Art. no. RG2002, doi: 10.1029/2011RG000372.

[20] A. Gruber, W. A. Dorigo, S. Zwieback, A. Xaver, and W. Wagner, "Characterizing coarse-scale representativeness of in situ soil moisture measurements from the international soil moisture network," *Vadose Zone J.*, vol. 12, no. 2, 2013, Art. no. vzj2012.0170.

[21] M. Z. Hakuba, D. Folini, A. Sanchez-Lorenzo, and M. Wild, "Spatial representativeness of ground-based solar radiation measurements," *J. Geophys. Res., Atmos.*, vol. 118, no. 15, pp. 8585–8597, 2013.

[22] G. Huang, X. Li, C. Huang, S. Liu, Y. Ma, and H. Chen, "Representativeness errors of point-scale ground-based solar radiation measurements in the validation of remote sensing products," *Remote Sens. Environ.*, vol. 181, pp. 198–206, 2016.

[23] Y. Ran *et al.*, "Spatial representativeness and uncertainty of eddy covariance carbon flux measurements for upscaling net ecosystem productivity to the grid scale," *Agricultural Forest Meteorol.*, vol. 230/231, pp. 114–127, 2016.

[24] A. Apte, M. Hairer, A. M. Stuart, and J. Voss, "Sampling the posterior: An approach to non-Gaussian data assimilation," *Phys. D, Nonlinear Phenomena*, vol. 230, no. 1, pp. 50–64, 2007.

[25] R. N. Miller, E. F. Carter, and S. T. Blue, "Data assimilation into nonlinear stochastic models," *Tellus A, Dyn. Meteorol. Oceanogr.*, vol. 51, no. 2, pp. 167–194, 1999.

[26] G. L. Eyink, J. M. Restrepo, and F. J. Alexander, "A mean field approximation in data assimilation for nonlinear dynamics," *Phys. D, Nonlinear Phenomena*, vol. 195, no. 3, pp. 347–368, 2004.

[27] R. N. Miller, "Topics in data assimilation: Stochastic processes," *Phys. D, Nonlinear Phenomena*, vol. 230, no. 1, pp. 17–26, 2007.

[28] F. Liu and X. Li, "Formulation of scale transformation in a stochastic data assimilation framework," *Nonliner Processes Geophys.*, vol. 24, no. 2, pp. 279–291, 2017.

[29] P. Billingsley, *Probability and Measure*, 2nd ed. New York, NY, USA: Wiley, 1986.

[30] R. G. Bartle, *The Elements of Integration and Lebesgue Measure*. New York, NY, USA: Wiley, 1995.

[31] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, 2nd ed. New York, NY, USA: Springer-Verlag, 1991.

[32] A. M. Fowler, S. L. Dance, and J. A. Waller, "On the interaction of observation and prior error correlations in data assimilation," *Quart. J. Roy. Meteorol. Soc.*, vol. 144, no. 710, pp. 48–62, 2018.

[33] F. Liu, L. Wang, X. Li, and C. Huang, "ComDA: A common software for nonlinear and non-Gaussian land data assimilation," *Environ. Model. Softw.*, vol. 127, 2020, Art. no. 104638.

[34] P. J. Sellers *et al.*, "A revised land surface parameterization (SiB2) for atmospheric GCMS. Part I: Model formulation," *J. Climate*, vol. 9, no. 4, pp. 676–705, 1996.

[35] P. J. Sellers, Y. Mintz, Y. C. Sud, and A. Dalcher, "A simple biosphere model (SIB) for use within general circulation models," *J. Atmos. Sci.*, vol. 43, no. 6, pp. 505–531, 1986.

[36] G. D. Colello, C. Grivet, P. J. Sellers, and J. A. Berry, "Modeling of energy, water, and CO_2 flux in a temperate grassland ecosystem with SiB2: May–October 1987," *J. Atmos. Sci.*, vol. 55, no. 7, pp. 1141–1169, Apr. 1998.

[37] N. P. Hanan *et al.*, "Testing a model of CO_2, water and energy exchange in Great Plains tallgrass prairie and wheat ecosystems," *Agricultural Forest Meteorol.*, vol. 131, no. 3, pp. 162–179, 2005.

[38] Z. Gao, N. Chae, J. Kim, J. Hong, T. Choi, and H. Lee, "Modeling of surface energy partitioning, surface temperature, and soil wetness in the Tibetan prairie using the Simple Biosphere Model 2 (SiB2)," *J. Geophys. Res., Atmos.*, vol. 109, no. D6, 2004, Art. no. D06102.

[39] C. Huang, X. Li, L. Lu, and J. Gu, "Experiments of one-dimensional soil moisture assimilation system based on ensemble Kalman filter," *Remote Sens. Environ.*, vol. 112, no. 3, pp. 888–900, 2008.

[40] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[41] X. Li *et al.*, "Heihe watershed allied telemetry experimental research (HiWATER): Scientific objectives and experimental design," *Bull. Amer. Meteorol. Soc.*, vol. 94, no. 8, pp. 1145–1160, Aug. 2013.

[42] R. Jin *et al.*, "A nested ecohydrological wireless sensor network for capturing the surface heterogeneity in the midstream areas of the Heihe River Basin, China," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 11, pp. 2015–2019, Nov. 2014.

[43] X. Han, R. Jin, X. Li, and S. Wang, "Soil moisture estimation using cosmic-ray soil moisture sensing at heterogeneous farmland," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 9, pp. 1659–1663, Sep. 2014.

[44] D. Li, R. Jin, J. Zhou, and J. Kang, "Analysis and reduction of the uncertainties in soil moisture estimation with the L-MEB model using EFAST and ensemble retrieval," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 6, pp. 1337–1341, Jun. 2015.

[45] X. Li *et al.*, "CASEarth poles: Big data for the three poles," *Bull. Amer. Meteorol. Soc.*, vol. 101, no. 9, pp. E1475–E1491, Sep. 2020, doi: 10.1175/bams-d-19-0280.1.

[46] M. Zreda *et al.*, "COSMOS: The COsmic-ray soil moisture observing system," *Hydrol. Earth Syst. Sci.*, vol. 16, no. 11, pp. 4079–4099, 2012, doi: 10.5194/hess-16-4079-2012.

[47] W. Tian, X. Li, G. D. Cheng, X. S. Wang, and B. X. Hu, "Coupling a groundwater model with a land surface model to improve water and energy cycle simulation," *Hydrol. Earth Syst. Sci.*, vol. 16, no. 12, pp. 4707–4723, 2012, doi: 10.5194/hess-16-4707-2012.

Feng Liu received the B.S. degree in information and computing sciences from the North China University of Technology, Beijing, China, in 2006, the M.S. degree in geographic information systems from the Graduate University of Chinese Academy of Sciences, Beijing, China, in 2013, and the Ph.D. degree in geographic information systems from the University of Chinese Academy of Sciences, Beijing, in 2017.

He is currently with the Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou, China. He is mainly engaged in interdisciplinary research work on mathematical modeling and data assimilation, spatial data analysis, and remote sensing.

Zebin Zhao received the B.S. degree in geographic information systems from Lanzhou Jiaotong University, Lanzhou, China, in 2014, and the M.S. and Ph.D. degrees in geographic information systems from the Chinese Academy of Sciences, Beijing, China, in 2017 and 2022, respectively.

He is currently a Postdoctoral Fellow with the Northwest Institute of Eco-Environment and Resources (originally the Cold and Arid Regions Environmental and Engineering Research Institute), Chinese Academy of Sciences, Lanzhou, China. His research interests include blind source separation, microwave remote sensing of soil moisture, and COVID-19 prediction.

Xin Li (Senior Member, IEEE) received the B.S. degree in geographic information systems from Nanjing University, Nanjing, China, in 1992, and the Ph.D. degree in geographic information systems from the Chinese Academy of Sciences (CAS), Beijing, China, in 1998.

Since 1999, he has been a Professor with the Cold and Arid Regions Environmental and Engineering Research Institute, CAS, Lanzhou, China. He is currently the Director and a Professor with the National Tibetan Plateau Data Center, Institute of Tibetan Plateau Research, CAS. He has led the Watershed Allied Telemetry Experimental Research and the Heihe Watershed Allied Telemetry Experimental Research, which are comprehensive remote sensing experiments conducted sequentially in recent years with more than 600 participants. His research interests include land data assimilation, the application of remote sensing and geography information system in hydrology and cryosphere science, and integrated watershed modeling.

Dr. Li is a Senior Member of the IEEE Geoscience and Remote Sensing Society, a member of the Global Energy and Water Exchanges Scientific Steering Committee, the International Science Advisory Panel of Global Water Futures, the American Geophysical Union, and the American Meteorological Society, and the Vice President of GEWEX China. He was the recipient of the several honors by CAS and the Chinese government for his outstanding contribution.