The Optimality of Potential Rescaling Approaches in Land Data Assimilation

M. TUGRUL YILMAZ AND WADE T. CROW

Hydrology and Remote Sensing Laboratory, Beltsville, Maryland

(Manuscript received 13 April 2012, in final form 28 September 2012)

ABSTRACT

It is well known that systematic differences exist between modeled and observed realizations of hydrological variables like soil moisture. Prior to data assimilation, these differences must be removed in order to obtain an optimal analysis. A number of rescaling approaches have been proposed for this purpose. These methods include rescaling techniques based on matching sampled temporal statistics, minimizing the least squares distance between observations and models, and the application of triple collocation. Here, the authors evaluate the optimality and relative performances of these rescaling methods both analytically and numerically and find that a triple collocation–based rescaling method results in an optimal solution, whereas variance matching and linear least squares regression approaches result in only approximations to this optimal solution.

1. Introduction

Given the availability of multiple approaches (i.e., models, in situ observations, and remote sensing) for estimating many geophysical variables, it is often desirable to merge them to obtain a more accurate product. In data assimilation, the goal is to optimally merge independent datasets with different error characteristics to obtain an analysis product with higher accuracy than all of the parent products.

However, the use of different modeling and/or observational approaches typically leads to predictions with different systematic relationships to the assumed truth. This is particularly true for soil moisture data assimilation given well-known climatological differences in both model-derived (Koster et al. 2009) and remotely sensed (Jackson et al. 2010) soil moisture products. Additionally, absolute values of models and observations differ from ground observations (Reichle and Koster 2004; Reichle et al. 2004). Hence, it is crucial to remove systematic differences between different datasets before using them in a hydrological data assimilation framework (Reichle and Koster 2004). This is commonly achieved by rescaling soil moisture observations to match model-predicted soil moisture (in some statistical sense) during a preprocessing step.

Several potential strategies for such rescaling have been proposed and applied in recent land data assimilation studies. Among them, cumulative distribution function (CDF) matching (Reichle and Koster 2004) and variance matching techniques are perhaps the most common. A handful of studies have applied rescaling based on least squares regression techniques (Crow et al. 2005; Crow and Zhan 2007) but failed to offer any clear rationale for this choice. Additionally, signal variance-based rescaling, typically applied as a preprocessing step in triple collocation analysis (Stoffelen 1998), also provides a means to rescale datasets using three independent estimates of the same variable. However, this approach has not yet been applied in soil moisture data assimilation.

Although there are many existing methods for rescaling hydrological variables, their optimality in terms of analysis errors in an assimilation framework has not yet been assessed. This paper investigates the relative performances of the above-mentioned rescaling methods both analytically and numerically.

The theoretical rationale for rescaling, and the degree to which rescaling techniques discussed above are consistent with this rationale, are discussed in the next section. Section 3 briefly presents the numerical experiment setup, section 4 presents the numerical results,

Corresponding author address: M. Tugrul Yilmaz, Hydrology and Remote Sensing Laboratory, Agricultural Research Service, U.S. Department of Agriculture, 10300 Baltimore Ave., BARC-WEST, Bldg. 007, Room 104, Beltsville, MD 20705.
E-mail: tugrul.yilmaz@ars.usda.gov

DOI: 10.1175/JHM-D-12-052.1

section 5 discusses the implications of the results, and section 6 summarizes our conclusions.

2. Rescaling datasets

a. Analytical solution for the rescaling factor

Assuming that we have representations of a given geophysical variable derived from both a model and observations, we can generalize the model-based estimates $\mathbf{x}$ and the observations $\mathbf{y}$ in a linear form as

$$\mathbf{x} = \mu_x + \alpha_x \mathbf{t}' + \boldsymbol{\epsilon}_x \quad \text{and} \tag{1}$$

$$\mathbf{y} = \mu_y + \alpha_y \mathbf{t}' + \boldsymbol{\epsilon}_y, \tag{2}$$

where μ_x and μ_y are the mean values of $\mathbf{x}$ and $\mathbf{y}$, $\mathbf{t}'$ is the true anomaly of the geophysical variable, α_x and α_y are scaling factors between the magnitude of the anomaly signals of $\mathbf{x}$ and $\mathbf{y}$ with $\mathbf{t}'$, and $\boldsymbol{\epsilon}_x$ and $\boldsymbol{\epsilon}_y$ are zero mean random errors in $\mathbf{x}$ and $\mathbf{y}$. In hydrological data assimilation, observations $\mathbf{y}$ are derived from in situ measurements and/or satellite-based retrievals, $\boldsymbol{\epsilon}_y$ is commonly assumed to lack autocorrelation, and $\boldsymbol{\epsilon}_x$ is generally considered to contain autocorrelation owing to the temporal memory of the model. In this setup, μ and $\alpha t'$ represent the signal component while ϵ represents the noise component. We emphasize that (1) and (2) are general enough to encompass any unit dimension and dynamic range differences that may exist between $\mathbf{x}$ and $\mathbf{y}$. In addition, note that, for the case in which the observations are assumed to capture a linear transformation of $\mathbf{t}'$ (rather than $\mathbf{t}'$ itself), the required transformation can simply be folded into the existing linear form of (2) through a trivial redefinition of α_y. As a result, the development below is equally valid for the case of a linear observation operator.

The purpose of data assimilation should be to reduce the magnitude of the noise component while preserving the information obtained from the signal components. Although these products have similarities in the way they realize the truth, they often have characteristic differences as well (i.e., different μ and α). Therefore, without the knowledge of the truth, arguably the best way to ensure the merged product has minimized error variance (assuming the uncertainties of products are characterized accurately) is to match datasets x and y to minimize the systematic differences between them prior to data assimilation. Without knowledge of the truth, matching datasets can be done by selecting one of the datasets as reference and linearly rescaling the other one.

Given the linear model in (1) and (2) and assuming x is the reference dataset, y can be rescaled via the general linear transformation

$$\mathbf{y}^* = \mu_x + (\mathbf{y} - \mu_y)c_y, \tag{3}$$

where c_y is a rescaling factor and y^* the rescaled dataset. Combining (2) and (3), we obtain

$$\mathbf{y}^* = \mu_x + \alpha_y c_y \mathbf{t}' + \boldsymbol{\epsilon}_y c_y. \tag{4}$$

After this rescaling, error in $\mathbf{y}^*$ can be expressed as

$$\boldsymbol{\epsilon}_{y^*} = \mathbf{y}^* - \mathbf{t} \quad \text{and} \tag{5}$$

$$\boldsymbol{\epsilon}_{y^*} = (\alpha_y c_y - \alpha_x)\mathbf{t}' + \boldsymbol{\epsilon}_y c_y, \tag{6}$$

where the unknown truth is given as $\mathbf{t} = \mu_x + \alpha_x \mathbf{t}'$ since $\mathbf{x}$ is the reference dataset. Our goal here is identifying the functional form of c_y that leads to an optimal data assimilation analysis. A key condition for such optimality is that assimilated observations $\mathbf{y}^*$ have orthogonal errors or $E[\mathbf{t}\boldsymbol{\epsilon}_{y^*}^{\mathrm{T}}] = 0$ (Chui and Chen 1998, p. 33). The orthogonality of $\boldsymbol{\epsilon}_{y^*}$ can expressed as

$$E[\mathbf{t}\boldsymbol{\epsilon}_{y^*}^{\mathrm{T}}] = E[(\mu_x + \alpha_x \mathbf{t}')\{(\alpha_y c_y - \alpha_x)\mathbf{t}' + \boldsymbol{\epsilon}_y c_y^T\}] \tag{7}$$

$$= \alpha_x(\alpha_y c_y - \alpha_x)\sigma_{t'}^2, \tag{8}$$

where the expectation operation $E[\cdot]$ represents long-term temporal averaging, $\sigma_{t'}^2 (E[\mathbf{t}'\mathbf{t}'^{\mathrm{T}}])$ is the variance of $\mathbf{t}'$, and $E[\mathbf{t}\boldsymbol{\epsilon}_{\mathbf{y}}^{\mathrm{T}}] = 0$ is assumed. Accordingly, in order for $\boldsymbol{\epsilon}_{y^*}$ to satisfy the orthogonal property, (8) has to vanish. The only nontrivial way to ensure this is defining the optimal value of c_y (c_y^O) as

$$c_y^O = \frac{\alpha_x}{\alpha_y}. \tag{9}$$

Note that (9) can also be obtained from optimal filtering requirements that errors in $\mathbf{y}^*$ be uncorrelated in time and/or errors in the analysis must be orthogonal.

b. Numerical solutions for the rescaling factor

Since α_x and α_y are typically unknown, (9) cannot be calculated directly. Instead, most land data assimilation studies attempt to replicate (9) using data that are available (i.e., $\mathbf{x}$ and $\mathbf{y}$). Therefore, it is useful to consider the relationship between functional forms of c_y derived from potential empirical rescaling strategies and the optimal form in (9). In appendix A we derive functional forms of c_y obtained 1) by using linear least squares techniques to regress $\mathbf{y}$ onto $\mathbf{x}$ (c_y^R), 2) by scaling y so that it has the same long-term temporal variance as $\mathbf{x}$ (c_y^V), and 3) by applying preprocessing techniques commonly used in triple collocation (c_y^T) as

$$c_y^R = \frac{\alpha_x}{\alpha_y}\frac{\alpha_y^2\sigma_{t'}^2}{\sigma_y^2}, \tag{10}$$

$$c_y^V = \frac{\alpha_x}{\alpha_y}\frac{(1+\sigma_{\epsilon_x}^2/\alpha_x^2\sigma_t^2)^{1/2}}{(1+\sigma_{\epsilon_y}^2/\alpha_y^2\sigma_t^2)^{1/2}}, \quad \text{and} \tag{11}$$

$$c_y^T = \frac{\alpha_x}{\alpha_y}, \tag{12}$$

where σ_y^2 is the variance of y, σ_t^2 is the variance of t (also note that $\sigma_{t'}^2 = \sigma_t^2$), and $\sigma_{\epsilon_x}^2$ and $\sigma_{\epsilon_y}^2$ are the error variances of **x** and **y**, respectively. Defining the signal variance ($\alpha_y^2\sigma_{t'}^2$ or $\alpha_x^2\sigma_{t'}^2$) to total variance ratio of **y** (str_y) and **x** (str_x) as

$$\text{str}_y = \frac{\alpha_y^2\sigma_{t'}^2}{\sigma_y^2} \quad \text{and} \tag{13}$$

$$\text{str}_x = \frac{\alpha_x^2\sigma_{t'}^2}{\sigma_x^2} \tag{14}$$

and using (9), (10)–(12) can be rewritten as

$$c_y^R = c_y^O \text{str}_y, \tag{15}$$

$$c_y^V = c_y^O\left(\frac{\text{str}_y}{\text{str}_x}\right)^{1/2}, \quad \text{and} \tag{16}$$

$$c_y^T = c_y^O, \tag{17}$$

where σ_x^2 is the variance of **x**. The expressions for c_y^R in (10) and c_y^V in (11) can also be written as

$$c_y^R = \frac{\sigma_x}{\sigma_y}\rho_{(x,y)} \quad \text{and} \tag{18}$$

$$c_y^V = \frac{\sigma_x}{\sigma_y}, \tag{19}$$

where $\rho_{(x,y)}$ is the correlation between **x** and **y** (see appendix A). In these forms (18) and (19), c_y^R and c_y^V can be obtained in any data assimilation study.

Note that, considering the definition in (4), c_y^V in (19) ensures only the variances of **x** and **y*** match. This is sufficient for linear systems with Gaussian errors. However, a more general form of matching is also common in which the higher-order statistical moments are also matched. These so-called CDF matching approaches will be considered in numerical examples presented below.

c. *Optimal versus suboptimal solutions*

Above we derive the optimal solution for c_y in a sequential filtering framework as $c_y^O = \alpha_x/\alpha_y$ (9). We also define c_y for three different empirical rescaling strategies (c_y^R, c_y^V, and c_y^T). Of the three approaches considered [REG, VAR, and TCA abbreviations will be used to refer to the least squares regression–, variance matching–, and triple collocation analysis–based approaches described above; ORG will be used to refer to the nonrescaled original observations case of $c_y = 1$], only the TCA-based approach resulted in the optimal solution, whereas REG- and VAR-based solutions resulted in approximations to this optimal solution c_y^O in the form $c_y^O f$. For the REG- and VAR-based solutions, f factors are defined as $f_R = \text{str}_y$ in (15) and $f_V = (\text{str}_y/\text{str}_x)^{1/2}$ in (16), respectively. In all cases where f_R or f_V are not equal to one, these two approaches diverge from the optimal solution given by (9). Therefore, the suboptimal REG-based solution converges to the optimal solution as str_y converges to one, while VAR-based solution converges to the optimal solution only when $\text{str}_x = \text{str}_y$. Given the ubiquity of VAR- and REG-based rescaling approaches in contemporary land data assimilation, this demonstrates that a widely applied element of existing assimilation systems is generally suboptimal. An optimal solution is available from TCA-based rescaling approach; however, it requires three independent and mutually linear datasets of sufficient temporal length. If these requirements are not met, which is generally the case for most hydrological data assimilation systems, we are limited to the approximate REG- and VAR-based solutions.

3. Synthetic-twin experiment setup

The failure of REG- and VAR-based preprocessing to generally conform to the optimal rescaling state criteria given in (9) should degrade the subsequent performance of a sequential filter. Here, we conduct a series of synthetic-twin data assimilation experiments to quantify this degradation. Our numerical results are based on a simple antecedent precipitation index (API) model, given as

$$x_d = \gamma_d x_{d-1} + P_d \quad \text{and} \tag{20}$$

$$\gamma_d = a + b\cos(2\pi d/365), \tag{21}$$

where d is day of the year, x_d is the API model value at d, P_d is the precipitation value at d, and a and b values are selected as 0.85 and 0.10, respectively. The model is run over a single 0.25° pixel (35°N, 98°W) using daily Tropical Rainfall Measuring Mission (TRMM) 3B42 precipitation accumulations acquired between 1998 and 2010.

TABLE 1. Standard deviation cases for random additive observation and model perturbations σ_{om} (20 cases total).

σ_{om}(1–10)	σ_{om}(11–20)
1.5	4.4
1.6	5.5
1.8	7
1.85	8
1.9	9
1.95	10
2	12
2.5	15
3	20
3.6	25

Using the above API model, we have created daily synthetic ground truth **t**. Control runs x are obtained from model runs that do not assimilate observations, while API values from d to $d + 1$ are additively perturbed with random numbers that have mean of zero and standard deviations given in Table 1. Original observations **y** are created by multiplying the truth with a constant (true observation scaling factors α_y) and then adding mean-zero random noise with the same standard deviations as the control run (Table 1). We later rescale **y** to **x** by using four different rescaling methods: VAR observations are created using (19), CDF observations are created by using the CDF-matching technique described by Reichle and Koster (2004), REG observations are created using (18), and TCA observations are created using (12). For the TCA-based rescaling, **z** values are created in an identical way to **y** but using a different random number sequence.

Perturbation standard deviations (Table 1) and α_y values [α_y = (0.12, 1.00, 2.50)] are selected to result in increasing (or decreasing) c_y and/or $\rho_{(x,y)}$. The true rescaling factors c_y^O are calculated by taking the ratio of α_x/α_y. Here, the values of α_y are given as input in the experiment design and therefore explicitly known. However, α_x is not known and instead calculated as

$$\alpha_x = \frac{E[\mathbf{t}'\mathbf{x}'^{\mathrm{T}}]}{\sigma_{t'}^2}, \tag{22}$$

where $\mathbf{x}'$ is the control run anomaly. Rescaled observations are later assimilated into (20) using an ensemble Kalman filter (EnKF) of the form

$$x_e^+ = x_e^- + K(y_e^* - x_e^-), \tag{23}$$

where

$$K = \frac{\sigma_{\epsilon_x}^2}{\sigma_{\epsilon_x}^2 + \sigma_{\epsilon_{y^*}}^2} \tag{24}$$

and e is the ensemble member number (total is 40); x_e^+ and x_e^- are analysis and forecast model values, respectively, for e at d; y_e^* is the synthetically created linearly rescaled observation for e at d; and K is the Kalman gain at d. Here we note that the methodology is general to any Kalman filter variant while our choice of EnKF is arbitrary. Ensembles of observations are created by perturbing the observations at any given time step with statistics consistent with the error variances used for the calculation of K. An ensemble of model replicates at any time step are created by adding mean-zero noise (standard deviations given Table 1) to model forecasts of **x**. Values of σ_{ϵ_x} are sampled from the background ensemble of $\mathbf{x}_d$ at d, while observation error standard deviations $\sigma_{\epsilon_{y^*}}$ are calculated using the rescaled sets as

$$\sigma_{\epsilon_{y^*}} = \{E[(\mathbf{y}^* - \mathbf{t})(\mathbf{y}^* - \mathbf{t})^{\mathrm{T}}]\}^{1/2}. \tag{25}$$

Hence, we assign perfect error variances to **y***.

Using this synthetic-twin framework, we investigate the impact of the REG, VAR, CDF, and TCA rescaling strategies on the accuracy of subsequent EnKF predictions by estimating the error standard deviation of EnKF analysis $\{\sigma_{\epsilon_m} = (E[\boldsymbol{\epsilon}_m\boldsymbol{\epsilon}_m^{\mathrm{T}}])^{1/2}\}$ and the EnKF analysis correlation with the truth $\rho_{(m,t)}$. In particular, we investigate these estimates as a function of $\rho_{(x,y)}$ since it differentiates the suboptimal REG- and VAR-based solutions [in (18) and (19)].

4. Results

Based on the synthetic-twin EnKF setup described above, we examined the performance of various rescaling strategies by selecting three c_y^O cases (corresponding to low, close to one, and high $\mathrm{str}_y/\mathrm{str}_x$) and using all 20 perturbation scenarios summarized in Table 1. Values of $\sigma_{\epsilon_{y^*}}$ for each of these three cases are plotted in separate panels in Fig. 1 (each line in each panel contains results for 20 different model and observation perturbation values). For the same three cases, σ_{ϵ_m} and $\rho_{(m,t)}$ are presented in Figs. 2 and 3 (similar to Fig. 1, different model and observation perturbation values are plotted for each rescaling method).

Confirming the earlier theoretical analysis, TCA-based rescaling results in the smallest σ_{ϵ_m} for all cases (Figs. 2a–c) with the exception when $\rho_{(x,y)}$ are very low. VAR-based σ_{ϵ_m} are very similar to the TCA-based σ_{ϵ_m} when $(\mathrm{str}_y/\mathrm{str}_x)^{0.5}$ (subscripts x and y refer to the model and observations, respectively) are around one (Fig. 2b). CDF-based σ_{ϵ_m} (not shown) are very similar to VAR-based σ_{ϵ_m} for all cases. Additionally, the impact of

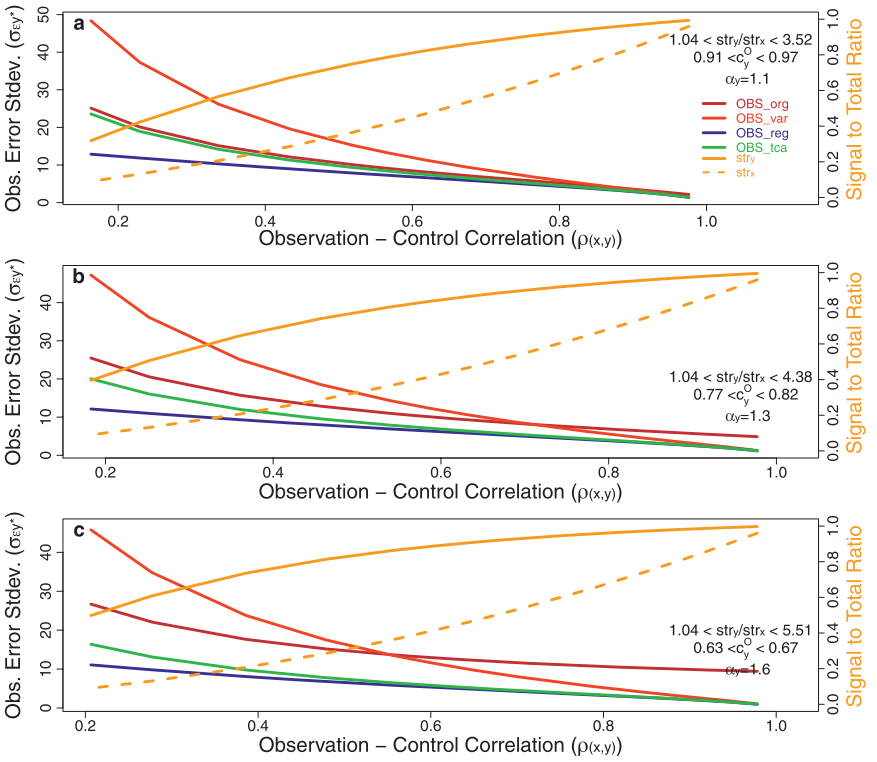

FIG. 1. Observation error standard deviations (left axis) and str of observations [see (13)] and the model [see (14)] (both on right axis), for (a) very low (0.12), (b) unity (1.00), and (c) very high (2.5) values of α_y.

rescaling approaches (i.e., the spread in VAR-, REG-, and TCA-based results) is maximized when $\rho_{(x,y)}$ are minimized, which emphasizes the importance of accurate rescaling for variables having moderate to low model/observation correlations (such as soil moisture).

Limited cases, where REG- and VAR-based rescaling produces smaller σ_{ϵ_m} than TCA-based rescaling, are attributed to the lack of reliability of error variance/standard deviation as a performance metric when comparing datasets with different dynamic ranges (Entekhabi et al. 2010; Gupta et al. 2009). In particular, appendix B demonstrates how suboptimal solutions (VAR and REG based) can produce spuriously low error variances when $c_y \gg 1$ and str_x and str_y are very low. This problem is especially acute for REG-based rescaling when $c_y \gg 1$ since it frequently results in rescaled datasets with very small standard deviations due to grossly underestimated rescaling factors when $\text{str}_y \ll 0.5$. Hence, it is necessary to replot Fig. 2 using $\rho_{(m,t)}$ as an alternative error metric.

Results in Fig. 3 demonstrate that TCA-based rescaling results in the highest $\rho_{(m,t)}$ for all examined cases. In addition, confirming earlier theoretical results, REG- and TCA-based EnKF results have comparable $\rho_{(m,t)}$ when str_y values are high (Fig. 3c), and VAR-based rescaling converges to TCA-based rescaling when str_x and str_y are approximately equal (Fig. 3b).

When α_y values are very low, TCA-based $\sigma_{\epsilon_{y^*}}$ are the highest (Fig. 1a), yet these observations result in the most accurate EnKF analysis (Fig. 3a). It can be shown that TCA-based $\sigma_{\epsilon_{y^*}}$ become higher than the errors of other rescaling approaches when $\text{str}_y < 0.5$ and $f < 1$ ($\text{str}_y/\text{str}_x < 1$ or $\text{str}_y < 1$). However, this does not imply anything wrong with TCA-based rescaling; on the contrary, it emphasizes the importance of correctly assigning rescaling factors and illustrates that the goal of rescaling is not necessarily to minimize $\sigma_{\epsilon_{y^*}}$ plotted in Fig. 1. Another interesting result in Fig. 2b and Fig. 3b is the suggestion that application of suboptimal rescaling methods can degrade the accuracy of the EnKF (relative to the ORG case of no rescaling) for cases in which little or no rescaling is required (i.e., $c_y^O \sim 1$). Consequently, blindly applying any suboptimal scaling method entails an element of risk.

5. Discussion

Given that we present two suboptimal solutions (REG- and VAR-based rescaling) that are widely applied in

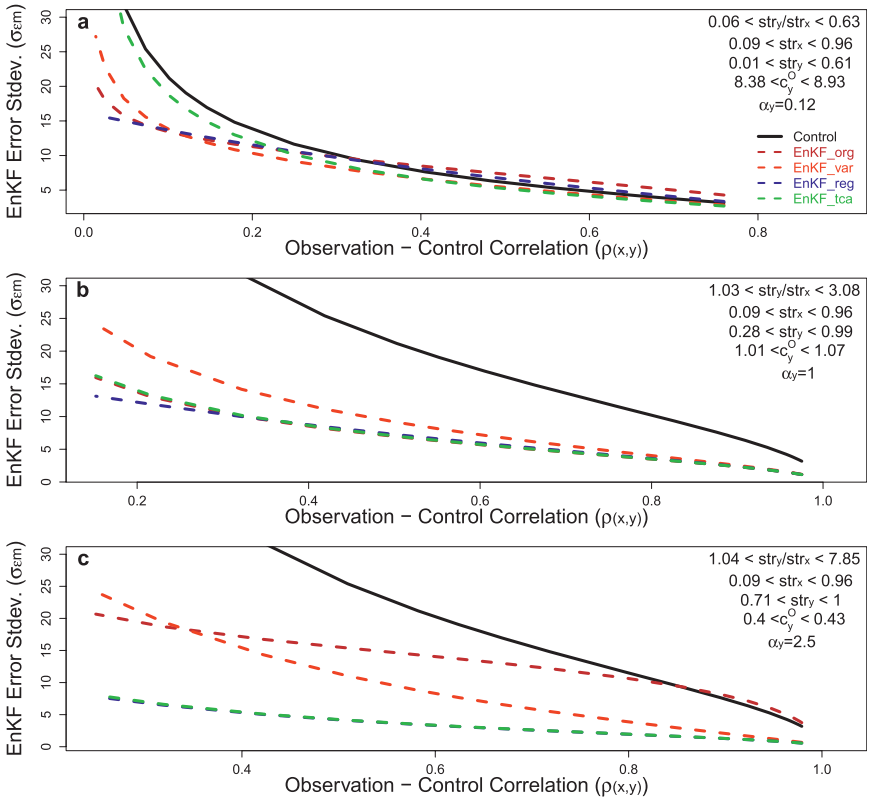

FIG. 2. EnKF analysis error standard deviations for three different α_y = (a) 0.12, (b) 1.00, and (c) 2.5. For clarity, actual str values (Fig. 1) are not drawn; instead, their max/min values are given. There are overlapping lines: green with brown in (b) and blue with green in (c).

hydrological sciences, it is of interest to generalize which one leads to a more accurate analysis under specific conditions. Theoretically, the relative accuracy of REG- and VAR-based rescaling depends on the relative magnitudes of str_y and $(\text{str}_y/\text{str}_x)^{0.5}$. However, such information is seldom readily available to developers of land data assimilation systems. Hence, it is not straightforward to offer general advice about whether the REG- or VAR-based rescaling method is optimal.

Nevertheless, it is possible to perform a consistency check to see whether a particular rescaling approach is consistent with statistical assumptions made during the implementation of a data assimilation system. For example, in the implementation of an EnKF, specific assumptions must be made regarding the error covariance of observations and the forecast uncertainty of the model. Based on these assumptions, estimates of str_x and str_y can be readily obtained [i.e., str_y and str_x can be found as $(\sigma_y^2 - \sigma_{\epsilon_y}^2)/\sigma_y^2$ and $(\sigma_x^2 - \sigma_{\epsilon_x}^2)/\sigma_x^2$, respectively]. Therefore, a consistency check is possible between these str estimates and rescaling methods. Excluding the special cases given in appendix B, if assumed $\text{str}_y/\text{str}_x \leq 1$ (observations less accurate then model or equally accurate), then VAR-based rescaling is preferable because the underestimation of rescaling factor through VAR-based rescaling would be less than through REG-based rescaling. Conversely, if observations are assumed to be more accurate than the model (i.e., $\text{str}_y/\text{str}_x > 1$), particularly when $\text{str}_y/\text{str}_x \gg 1$ and/or str_y is high ($\gg$0.4), then REG-based rescaling is preferable. In general, the choice of REG- or VAR-based rescaling methods is less critical (perhaps negligible) for very high str_y and str_x values (str > 0.9). However, note that particular str thresholds acquired from Fig. 2 (e.g., 0.4 and 0.9) might be system specific and not generalizable to other assimilation setups using different land models and/or observations. Nevertheless, at a minimum, this consistency check ensures that an applied rescaling approach is not grossly inconsistent with the error assumptions underlying the application of a particular data assimilation approach.

Another important issue is the relevance of this analysis for the case of utilizing an observation operator to directly assimilate satellite brightness temperature T_b observations rather than geophysical retrievals based on the inversion of T_b. One interesting implication of

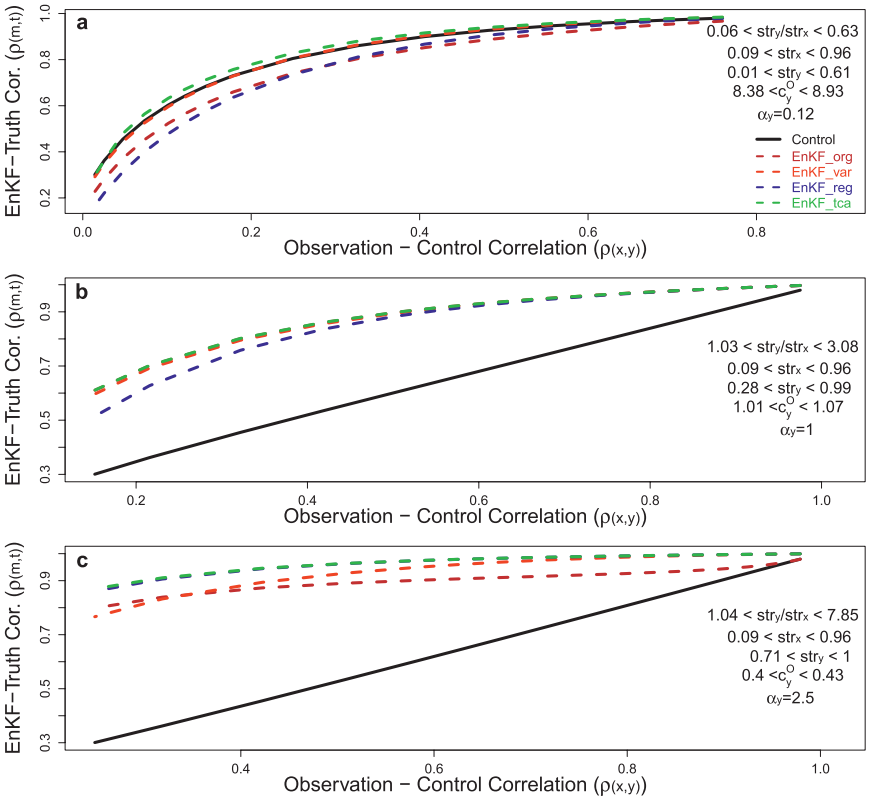

FIG. 3. As in Fig. 2, except EnKF analysis correlations with truth are plotted on the left axis. Actual str values are shown in Fig. 1. There are overlapping lines: brown and green in (b) and blue with green in (c).

applying a forward model to assimilate T_b is that the errors due to the radiance transfer model are effectively moved from the observation side to the model forecast side of the data assimilation system. As a consequence, assimilating T_b rather than soil moisture leads to an effective decrease in model-based str (str_x) and increase in observation str (str_y). In many cases, str_y could be quite close to one, since the accuracy goal of low-frequency (<10 GHz) satellite T_b retrievals used for soil moisture retrieval (often on the order of 1–3 K) tends to be small relative to the observation dynamic range in true T_b (up to 100 K). This suggests that a REG-based rescaling approach is advantageous for rescaling T_b observations prior to their assimilation as it yields smaller analysis errors when str_y is high and $\text{str}_y > \text{str}_x$ (Fig. 2). However, it should be stressed that, while results presented here can be trivially generalized for the application of a linear observation operator, it is currently unknown how significantly they are impacted by the presence of a strongly nonlinear observation operator. Therefore, additional analysis will be required to fully describe the implications of this analysis for T_b assimilation based on nonlinear forward radiative transfer calculations.

6. Conclusions

In hydrological assimilation studies, the primary goal is to combine different datasets to obtain a more accurate one via reducing the level of noise in the datasets. However, if datasets do not have a similar systematic relationship with the assumed truth, merging methodologies can result in increased errors even if the product uncertainties are specified correctly. As a result, it is critical to have correctly rescaled datasets before a merging methodology is applied.

This paper investigated existing methods that are widely applied in hydrological data assimilation studies to rescale observations prior to their assimilation into models. Specifically, we have evaluated the VAR-, CDF-, REG-, and TCA-based rescaling methods. Among these methods, the REG-based linear regression solution has been recognized by some studies (Gupta et al. 2009; Holmes et al. 2012) and applied by Crow et al. (2005) and Crow and Zhan (2007), whereas the vast majority of the hydrological assimilation studies have applied VAR- and CDF-based rescaling strategies. Although the triple collocation solution of Stoffelen (1998) has been widely

applied, it was not particularly emphasized before that its intermediate rescaling step should be applied in hydrological data assimilation studies.

In a hydrological assimilation study, if the errors of the reference and the matched datasets (i.e., hydrological model and the observations) are assumed negligible when compared to the real signal (implying very high str values), then these suboptimal rescaling factor solutions give very close to optimal estimates. However, for many hydrological studies the noise of the datasets cannot be ignored, hence the rescaling method should also take into account the magnitude of the noise components of both datasets. Among the methods, VAR- and CDF-based rescaling methods match the total variance of observations to the model while neglecting the noise contributions of the datasets (Gao et al. 2007), whereas the REG-based rescaling takes into account these error components via the additional multiplication factor of the correlation coefficient. Nevertheless, the VAR-, CDF-, and REG-based rescaling methods are only suboptimal solutions as they generally violate the orthogonality property of an optimal estimation procedure (section 2a). As a result, they provide only approximations to the optimal estimate with a multiplication factor f ($f_R = \mathrm{str}_y$ for the REG-based solution and $f_V = (\mathrm{str}_y/\mathrm{str}_x)^{1/2}$ for the VAR-based solution). Hence, the suboptimality of these methods is reduced and their solutions converge to the true solution only when f_R or f_V converge to one.

This analytical description of c_y^O is confirmed via a set of numerical synthetic-twin experiments using a simple soil moisture model. After rescaling the observations with the VAR-, CDF-, REG-, and TCA-based methods, we find that TCA rescaling leads to the most accurate EnKF analysis—with the exception of the cases clarified in appendix B. Accordingly, it is best to use TCA-based rescaling factors when available as long as its underlying assumptions (independence of errors, mutual linear relation, and long enough datasets) are met. If these conditions cannot be met, which often is the case in hydrological assimilation studies, then suboptimal approximations (VAR-, CDF-, and REG-based rescaling) must be used. However, in such cases, it should be recognized that such rescaling introduces a suboptimal element into the analysis that may degrade subsequent data assimilation results. The relative optimality of these approximations depends on the str of model and observations. Therefore, developers of land data assimilation applying a particular suboptimal approach should examine the consistency of their rescaling approach with error assumptions underlying the application of their data assimilation system. While a simple model is sufficient to clarify the underlying rescaling principles described above, follow-on work with more complex land surface models will be required to fully quantify the overall impact of rescaling errors on land data assimilation analysis products.

Acknowledgments. We thank two anonymous reviewers and Bart Forman for their constructive comments, which led to numerous clarifications in the final version of the manuscript. Research was partially supported by Wade Crow's membership in the NASA Soil Moisture Active/Passive Science Definition Team. The United States Department of Agriculture is an equal opportunity provider and employer.

APPENDIX A

Numerical Solutions for the Rescaling Factor

a. Rescaling factor from linear least squares regression

One potential rescaling strategy is choosing a value of c_y by linearly regressing $\mathbf{y}$ onto $\mathbf{x}$ and obtaining the best linear expression for $\mathbf{x}$ in terms of $\mathbf{y}$. This least squares sense solution can be found by minimizing the mean square difference (msd) between $\mathbf{x}$ and $\mathbf{y}^*$:

$$\mathrm{msd} = E[(\mathbf{x} - \mathbf{y}^*)^2]. \tag{A1}$$

Using (1) and (3) above can be written as

$$\mathrm{msd} = E[\{(\mu_x + \alpha_x \mathbf{t}' + \boldsymbol{\epsilon}_x) - (\mu_x + \alpha_y c_y \mathbf{t}' + \boldsymbol{\epsilon}_y c_y)\}^2], \tag{A2}$$

$$\mathrm{msd} = E[(\alpha_x \mathbf{t}' + \boldsymbol{\epsilon}_x)^2 + (\alpha_y \mathbf{t}' + \boldsymbol{\epsilon}_y)^2 c_y^2 - 2(\alpha_x \mathbf{t}' + \boldsymbol{\epsilon}_x)(\alpha_y \mathbf{t}' + \boldsymbol{\epsilon}_y) c_y], \quad \text{and} \tag{A3}$$

$$\mathrm{msd} = \sigma_x^2 + c_y^2 \sigma_y^2 - 2 c_y \alpha_x \alpha_y \sigma_{t'}^2. \tag{A4}$$

By taking the first derivative of (A4) with respect to c_y and setting it to zero, we find the regression-based rescaling factor solution c_y^R

$$\frac{\partial \mathrm{msd}}{\partial c_y^R} = 2 c_y^R \sigma_y^2 - 2 \alpha_x \alpha_y \sigma_t^2 = 0, \tag{A5}$$

$$c_y^R = \frac{\alpha_x \alpha_y \sigma_{t'}^2}{\sigma_y^2}, \quad \text{and} \tag{A6}$$

$$c_y^R = \frac{\alpha_x}{\alpha_y} \frac{\alpha_y^2 \sigma_{t'}^2}{\sigma_y^2}. \tag{A7}$$

Using the definitions of str_y in (13) and str_x in (14), c_y^R in (A7) can be written as

$$c_y^R = \frac{\alpha_x}{\alpha_y}\mathrm{str}_y. \tag{A8}$$

However, the right-hand side of (A8) cannot always be obtained in this form since the calculation of α_x, α_y, and str_y requires additional ground truth or ancillary datasets that are often not available. Consequently, we will rewrite (A8) in terms of readily available variables. To do this, we apply the definition of correlation between model and observation $\rho_{(x,y)}$:

$$\rho_{(x,y)} = \frac{E[(\mathbf{x}-\mu_x)(\mathbf{y}-\mu_y)]}{\sigma_x\sigma_y} \quad \text{and} \tag{A9}$$

$$\rho_{(x,y)} = \frac{E[(\alpha_x\mathbf{t}'+\boldsymbol{\epsilon}_x)(\alpha_y\mathbf{t}'+\boldsymbol{\epsilon}_y)]}{\sigma_x\sigma_y}. \tag{A10}$$

Assuming $E[\mathbf{t}'\boldsymbol{\epsilon}_x^{\mathrm{T}}]=0$, $E[\mathbf{t}'\boldsymbol{\epsilon}_y^{\mathrm{T}}]=0$, and $E[\boldsymbol{\epsilon}_y\boldsymbol{\epsilon}_x^{\mathrm{T}}]=0$, (A10) can be written as

$$\rho_{(x,y)} = \frac{\alpha_x\alpha_y\sigma_{t'}^2}{\sigma_x\sigma_y}, \tag{A11}$$

$$\rho_{(x,y)} = \frac{\alpha_x}{\alpha_y}\frac{\alpha_y^2\sigma_{t'}^2}{\sigma_y^2}\frac{\sigma_y}{\sigma_x}, \quad \text{and} \tag{A12}$$

$$\frac{\sigma_x}{\sigma_y}\rho_{(x,y)} = \frac{\alpha_x}{\alpha_y}\mathrm{str}_y. \tag{A13}$$

Using (A13), the regression-based rescaling factor solution in (A8) can be rewritten as

$$c_y^R = \frac{\sigma_x}{\sigma_y}\rho_{(x,y)}. \tag{A14}$$

In this form, c_y^R can be obtained in any data assimilation application.

b. Rescaling factor from variance matching

A widely applied rescaling strategy is transforming **y** so that its statistical moments match that of **x**. Since the form of (3) already ensures a match in means, the simplest viable case of this transformation is based solely on matching variances. Here, the rescaling factor from variance matching c_y^V is given by

$$c_y^V = \frac{\sigma_x}{\sigma_y}. \tag{A15}$$

Below we rewrite (A15) in a way that resembles (A8) to clarify later the differences between different rescaling factor solutions. Assuming the noise components of the datasets are independent from the truth ($E[\mathbf{t}\boldsymbol{\epsilon}_x^{\mathrm{T}}]=0$ and $E[\mathbf{t}\boldsymbol{\epsilon}_y^{\mathrm{T}}]=0$), σ_x^2 and σ_y^2 can be written as

$$\sigma_x^2 = \alpha_x^2\sigma_t^2 + \sigma_{\epsilon_x}^2 \quad \text{and} \tag{A16}$$

$$\sigma_y^2 = \alpha_y^2\sigma_t^2 + \sigma_{\epsilon_y}^2. \tag{A17}$$

Using these definitions, (A15) becomes

$$c_y^V = \frac{(\alpha_x^2\sigma_t^2+\sigma_{\epsilon_x}^2)^{1/2}}{(\alpha_y^2\sigma_t^2+\sigma_{\epsilon_y}^2)^{1/2}}, \tag{A18}$$

$$c_y^V = \frac{\alpha_x}{\alpha_y}\frac{(1+\sigma_{\epsilon_x}^2/\alpha_x^2\sigma_t^2)^{1/2}}{(1+\sigma_{\epsilon_y}^2/\alpha_y^2\sigma_t^2)^{1/2}}, \quad \text{and} \tag{A19}$$

$$c_y^V = \frac{\alpha_x}{\alpha_y}\left(\frac{\mathrm{str}_y}{\mathrm{str}_x}\right)^{1/2}. \tag{A20}$$

c. Rescaling factor from triple collocation

Triple collocation analysis (TCA) is an error magnitude estimation method that uses three linearly related independent products to obtain the errors of each product separately. It was initially introduced for error magnitude estimation in oceanic studies (Stoffelen 1998; Caires and Sterl 2003), and has recently been applied to large-scale soil moisture error estimation-based studies (Scipal et al. 2008; Parinussa et al. 2011; Hain et al. 2011; Yilmaz et al. 2012; Anderson et al. 2012). These studies are typically based on one model-based soil moisture product and two remotely sensed products derived from contrasting remote sensing retrieval techniques (e.g., passive and active microwave).

As a first step, TCA requires products to be rescaled to each other to match their signal statistics. Following Stoffelen (1998), such rescaling is based on applying TCA-based rescaling factor c_y^T

$$c_y^T = \frac{\overline{x'z'}}{\overline{y'z'}} \tag{A21}$$

in (3), where **z** is a third independent product that is similar to **x** and **y** (1)–(2), and defined as $\mathbf{z} = \mu_z + \alpha_z\mathbf{t}' + \boldsymbol{\epsilon}_z$ with time anomaly $\mathbf{z}' = \alpha_z\mathbf{t}' = \boldsymbol{\epsilon}_z$. Assuming all product errors are independent from both the truth and each other, (A21) can be rewritten as

$$c_y^T = \frac{\overline{(\alpha_x \mathbf{t}' + \boldsymbol{\epsilon}_x)(\alpha_z \mathbf{t}' + \boldsymbol{\epsilon}_z)}}{\overline{(\alpha_y \mathbf{t}' + \boldsymbol{\epsilon}_y)(\alpha_z \mathbf{t}' + \boldsymbol{\epsilon}_z)}}, \tag{A22}$$

$$c_y^T = \frac{\alpha_x \alpha_z \sigma_{t'}^2}{\alpha_y \alpha_z \sigma_{t'}^2}, \quad \text{and} \tag{A23}$$

$$c_y^T = c_y^O. \tag{A24}$$

Therefore, by taking advantage of a third independent product, TCA is able to recover the optimal rescaling factor defined in (9).

APPENDIX B

Optimal versus Suboptimal Rescaling Error Variances

The difference between the error variances of the optimal (TCA) rescaling and suboptimal (VAR and REG) rescaling is compared. Assuming $E[\mathbf{t}\boldsymbol{\epsilon}_x^{\mathrm{T}}] = 0$, $E[\mathbf{t}\boldsymbol{\epsilon}_y^{\mathrm{T}}] = 0$, and $E[\boldsymbol{\epsilon}_y\boldsymbol{\epsilon}_x^{\mathrm{T}}] = 0$, following (6), the error variance of the optimal and the suboptimal analysis are found as

$$\sigma_{\text{opt}}^2 = (\alpha_y c_o - \alpha_x)^2 w_o^2 \sigma_t^2 + [(1 - w_o)^2 \sigma_{\epsilon_x}^2 + w_o^2 c_o^2 \sigma_{\epsilon_y}^2] \quad \text{and} \tag{B1}$$

$$\sigma_{\text{sub}}^2 = (\alpha_y c_s - \alpha_x)^2 w_s^2 \sigma_t^2 + [(1 - w_s)^2 \sigma_{\epsilon_x}^2 + w_s^2 c_s^2 \sigma_{\epsilon_y}^2], \tag{B2}$$

where c_o (c_y^O or c_y^T in section 2) and c_s (c_y^V or c_y^R in section 2) are the rescaling factors of the optimal and suboptimal rescaling factors, respectively, and w_o and w_s are the weights of the rescaled observations associated with the optimal and suboptimal rescaling factors, respectively. Given optimal analysis satisfies $\alpha_y c_o - \alpha_x = 0$, (B1) can be written as

$$\sigma_{\text{opt}}^2 = (1 - w_o)^2 \sigma_{\epsilon_x}^2 + w_o^2 c_o^2 \sigma_{\epsilon_y}^2. \tag{B3}$$

We are interested in relative magnitudes of optimal and suboptimal error variances, hence we investigate their difference using (B2) and (B3):

$$\sigma_{\text{opt}}^2 - \sigma_{\text{sub}}^2 = [(1 - w_o)^2 - (1 - w_s)^2]\sigma_{\epsilon_x}^2 + (w_o^2 c_o^2 - w_s^2 c_s^2)\sigma_{\epsilon_y}^2 - (\alpha_y c_s - \alpha_x)^2 w_s^2 \sigma_t^2, \quad \text{and} \tag{B4}$$

$$= (w_o^2 - 2w_o + 2w_s - w_s^2)\sigma_{\epsilon_x}^2 + (w_o^2 c_o^2 - w_s^2 c_s^2)\sigma_{\epsilon_y}^2 - (c_s - c_o)^2 \alpha_y^2 w_s^2 \sigma_t^2. \tag{B5}$$

We are particularly interested in this difference when str_x, str_y, and α_y are very low and $\text{str}_y < \text{str}_x$ [setup in Fig. 2a for very low $\rho_{(x,y)}$]. Furthermore, for the scenario in Fig. 2a $\alpha_y \ll 1$, hence $c_o \gg 1$.

For the VAR-based solution, $\sigma_{\epsilon_x}^2 \sim c_s^2 \sigma_{\epsilon_y}^2 \gg \sigma_t^2$ (considering both str_x and str_y are very low), hence $w_s \sim 0.5$ [considering $w_s = \sigma_{\epsilon_x}^2/(\sigma_{\epsilon_x}^2 + c_s^2 \sigma_{\epsilon_y}^2)$]. Since $\text{str}_y < \text{str}_x < 1$, then $f < 1$, hence $c_s < c_o$ (reminder $c_s = c_o f$). Accordingly, $c_s^2 \sigma_{\epsilon_y}^2 \ll c_o^2 \sigma_{\epsilon_y}^2$ (given $c_o \gg 1$), and $w_s \gg w_o$. Under this condition ($w_s \gg w_o$), the first term in (B5) can be approximated to $w_s(2 - w_s)\sigma_{\epsilon_x}^2$, which is approximately $0.75\sigma_{\epsilon_x}^2$ (considering $w_s \sim 0.5$). Since $c_s < c_o$, the assumption of $c_s \ll c_o$ overestimates the third term in (B5). Thus, this assumption overall results in a higher number (approximately $0.25\sigma_t^2$ since $\alpha_y^2 c_o^2 = \alpha_x^2 \sim 1$) to be subtracted in (B5). Using these approximations for the first and the third terms, (B5) is rewritten as

$$\sigma_{\text{opt}}^2 - \sigma_{\text{sub}}^2 \sim 0.75\sigma_{\epsilon_x}^2 + (w_o^2 c_o^2 - w_s^2 c_s^2)\sigma_{\epsilon_y}^2 - 0.25\sigma_t^2. \tag{B6}$$

Following above ($\sigma_{\epsilon_x}^2 \gg \sigma_t^2$), third term can be dropped and (B6) becomes

$$\sigma_{\text{opt}}^2 - \sigma_{\text{sub}}^2 \sim 0.75\sigma_{\epsilon_x}^2 + (w_o^2 c_o^2 - w_s^2 c_s^2)\sigma_{\epsilon_y}^2. \tag{B7}$$

Here $c_s < c_o$ and $w_o \ll w_s$, hence $w_s^2 c_s^2$ is slightly higher than $w_o^2 c_o^2$, which results in the second term in (B6) being negative. However, this term is still much lower than the first term ($\sim 0.75\sigma_{\epsilon_x}^2$), hence the overall $\sigma_{\text{opt}}^2 - \sigma_{\text{sub}}^2$ difference remains positive, which implies $\sigma_{\text{opt}}^2 > \sigma_{\text{sub}}^2$ for VAR solution.

Similarly, for the REG-based solution, $\text{str}_y \ll 1$, hence $c_s \ll c_o$. It follows that $c_s^2 \sigma_{\epsilon y}^2 \ll c_o^2 \sigma_{\epsilon y}^2$, hence $w_s \gg w_o$, which also results in the $\sigma_{\text{opt}}^2 - \sigma_{\text{sub}}^2$ difference being positive, parallel to the VAR-based solution. Accordingly, suboptimal rescaling solutions (VAR and REG based) can result in smaller error variance than optimal solutions (TCA based) when $\alpha_y \ll 1$, and str_x and str_y are very low. For these conditions, the REG-based rescaling strategy is particularly prone to spuriously low error variances since c_y^R tends to be lower than c_y^V, while both c_y^R and c_y^V underestimate c_y (given $c_y^V = (\text{str}_y/\text{str}_x)c_y$ and $c_y^R = \text{str}_y c_y$, if $\text{str}_y < \text{str}_x \ll 1$ then $c_y^R \ll c_y^V < c_y$).

REFERENCES

Anderson, W. B., B. F. Zaitchik, C. R. Hain, M. C. Anderson, M. T. Yilmaz, J. Mecikalski, and L. Schultz, 2012: Towards an

integrated soil moisture drought monitor for East Africa. *Hydrol. Earth Syst. Sci.,* **9,** 4587–4631.

Caires, S., and A. Sterl, 2003: Validation of ocean wind and wave data using triple collocation. *J. Geophys. Res.,* **108,** 3098, doi:10.1029/2002JC001491.

Chui, C. K., and G. Chen, 1998: *Kalman Filtering with Real-Time Applications.* Springer, 230 pp.

Crow, W. T., and X. Zhan, 2007: Continental-scale evaluation of remotely sensed soil moisture products. *IEEE Geosci. Remote Sens. Lett.,* **4,** 451–455.

——, R. Bindlish, and T. J. Jackson, 2005: The added value of spaceborne passive microwave soil moisture retrievals for forecasting rainfall-runoff partitioning. *J. Geophys. Res.,* **32,** L18401, doi:10.1029/2005GL023543.

Entekhabi, D., and Coauthors, 2010: The Soil Moisture Active Passive (SMAP) Mission. *Proc. IEEE,* **98,** 704–716.

Gao, H., E. F. Wood, M. Drusch, and M. F. Mccabe, 2007: Copula-derived observation operators for assimilating TMI and AMSR-E retrieved soil moisture into land surface models. *J. Hydrometeor.,* **8,** 413–429.

Gupta, H. V., H. Kling, K. K. Yilmaz, and G. F. Martinez, 2009: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *J. Hydrol.,* **377** (1–2), 80–91.

Hain, C. R., W. T. Crow, J. R. Mecikalski, M. C. Anderson, and T. Holmes, 2011: An intercomparison of available soil moisture estimates from thermal infrared and passive microwave remote sensing and land surface modeling. *J. Geophys. Res.,* **116,** D15107, doi:10.1029/2011JD015633.

Holmes, T. R. H., T. J. Jackson, R. H. Reichle, and J. B. Basara, 2012: An assessment of surface soil temperature products from numerical weather prediction models using ground-based measurements. *Water Resour. Res.,* **48,** W02531, doi:10.1029/2011WR010538.

Jackson, T. J., and Coauthors, 2010: Validation of Advanced Microwave Scanning Radiometer soil moisture products. *IEEE Trans. Geosci. Remote Sens.,* **48,** 4256–4272.

Koster, R. D., Z. Guo, R. Yang, P. A. Dirmeyer, K. Mitchell, and M. J. Puma, 2009: On the nature of soil moisture in land surface models. *J. Climate,* **22,** 4322–4335.

Parinussa, R. M., T. R. H. Holmes, M. T. Yilmaz, and W. T. Crow, 2011: The impact of land surface temperature on soil moisture anomaly detection from passive microwave observations. *Hydrol. Earth Syst. Sci.,* **15,** 3135–3151.

Reichle, R. H., and R. D. Koster, 2004: Bias reduction in short records of satellite soil moisture. *Geophys. Res. Lett.,* **31,** L19501, doi:10.1029/2004GL020938.

——, ——, J. Dong, and A. A. Berg, 2004: Global soil moisture from satellite observations, land surface models, and ground data: Implications for data assimilation. *J. Hydrometeor.,* **5,** 430–442.

Scipal, K., T. Holmes, R. de Jeu, V. Naeimi, and W. Wagner, 2008: A possible solution for the problem of estimating the error structure of global soil moisture data sets. *Geophys. Res. Lett.,* **35,** L24403, doi:10.1029/2008GL035599.

Stoffelen, A., 1998: Toward the true near-surface wind speed: Error modeling and calibration using triple collocation. *J. Geophys. Res.,* **103** (C4), 7755–7766.

Yilmaz, M. T., W. T. Crow, M. C. Anderson, and C. Hain, 2012: An objective methodology for merging satellite- and model-based soil moisture products. *Water Resour. Res.,* **48,** W11502, doi:10.1029/2011WR011682.