Water Resources Research

RESEARCH ARTICLE

Key Points:
- Define efficiency of data assimilation from an information theory perspective
- Measures the total information available to data assimilation versus the amount extracted by an (imperfect) parametric DA algorithm
- Application example is an application of the EnKF to soil moisture assimilation

Correspondence to:
G. Nearing,
gsnearing@ua.edu

The Efficiency of Data Assimilation

Grey Nearing[1] (ID), **Soni Yatheendradas[2,3]** (ID), **Wade Crow[4]** (ID), **Xiwu Zhan[5]** (ID), **Jicheng Liu[3,5]** (ID), **and Fan Chen[4]** (ID)

[1]Department of Geological Sciences, University of Alabama, Tuscaloosa, AL, USA, [2]Hydrologic Sciences Laboratory, NASA GSFC, Greenbelt, MD, USA, [3]ESSIC, University of Maryland, College Park, MD, USA, [4]Hydrology and Remote Sensing Laboratory, USDA-ARS, Beltsville, MD, USA, [5]NOAA NESDIS Center for Satellite Applications and Research, College Park, MD, USA

Abstract Data assimilation is the application of Bayes' theorem to condition the states of a dynamical systems model on observations. Any real-world application of Bayes' theorem is approximate, and therefore, we cannot expect that data assimilation will preserve all of the information available from models and observations. We outline a framework for measuring information in models, observations, and evaluation data in a way that allows us to quantify information loss during (necessarily imperfect) data assimilation. This facilitates quantitative analysis of trade-offs between improving (usually expensive) remote sensing observing systems versus improving data assimilation design and implementation. We demonstrate this methodology on a previously published application of the ensemble Kalman filter used to assimilate remote sensing soil moisture retrievals from Advanced Microwave Scattering Radiometer for Earth (AMSR-E) into the Noah land surface model.

1. Introduction

Data assimilation (DA) is one of the most common methods for extracting information from Earth-observing remote sensing observations or retrievals. One especially important example of this in the hydrologic sciences is soil moisture DA (De Lannoy et al., 2016). For instance, the NASA Soil Moisture Active/Passive (SMAP) mission (Entekhabi et al., 2010) offers a DA product as part of the baseline mission (Reichle et al., 2016). At least 21 of the 55 groups participating in the SMAP Early Adopter Program (Moran et al., 2015) use the ensemble Kalman filter (EnKF; Evensen, 2003) as a primary method for extracting information from SMAP data products. Soil moisture DA is used routinely in Land Data Assimilation Systems (LDAS; Kumar et al., 2008; McNally et al., 2017; Rodell et al., 2004; Xia et al., 2011) for hydrological and hydrometeorological modeling (Maggioni & Houser, 2017), as well as in many other hydrology-related remote sensing applications (Mladenova et al., 2017).

Developing and deploying remote sensing instruments is expensive. Given that no DA algorithm or application will ever be able to perfectly extract the full information content of any remote sensing observations, we would like to have the ability to quantify how much information is present in a particular set of assimilation data versus the amount of information we are able to extract from those data using necessarily imperfect DA techniques. Maggioni et al. (2011) showed that overly simplistic precipitation uncertainty distributions degrade soil moisture DA results, and Maggioni et al. (2013) reported, more generally, that "a key issue of data assimilation is that observational and modeling uncertainties are poorly known, and incorrect assumptions about these errors may compromise (DA) efficiency." Another issue that is frequently encountered when evaluating DA results is that no in situ data measure exactly the same quantities that the model simulates or that the remote sensing platform observes. Our purpose here is to propose a general quantitative theory for measuring the effects of DA assumptions on overall DA performance in the presence of only imperfect evaluation data.

DA is defined in this paper as the application of Bayes' theorem to update the states of a dynamical simulation model by probabilistic conditioning on observation data. We will use the word *retrievals* to refer to the assimilated observation data (often, assimilated data are from satellites or other remote sensing platforms), and distinguish these from *evaluation data*, which are (typically in situ) observation data used to evaluate a particular DA application. However, notice that the theory we propose here is applicable to any type of assimilated observation data, not just satellite retrievals.

There are four sources of uncertainty in any DA application:

1. error in the dynamical systems model,
2. error in the assimilated retrievals,
3. approximations and assumptions in the DA algorithm itself, and
4. error in evaluation data.

The purpose of DA is to help mitigate the first two sources of uncertainty, by using imperfect retrievals to correct imperfect model states. But no DA system can mitigate model and retrieval error completely. In principle, there is a finite quantity of information about the true state of a dynamical system contained in the simulated state of an (imperfect) model, and there is also a finite quantity of information about the true state of that dynamical system in the (imperfect) retrievals. Taken together, there is some total quantity of information available to the DA algorithm from both model and retrievals together. However, because of errors and uncertainty in the model and errors and uncertainty in the retrievals, the total available information from both together is generally less than what is necessary to achieve perfectly accurate predictions of the real system states.

Related to the third source of uncertainty, we can imagine a hypothetically perfect DA algorithm that could extract and combine all of the incomplete information from both model and retrievals. In order to achieve perfect information extraction, this hypothetically perfect DA algorithm would necessarily include perfect (nonparametric) model and retrieval error distributions, would include a perfect (typically nonlinear) observation operator, and would be able to sample the Bayesian posterior with at least asymptotic efficiency (e.g., by Markov chain Monte Carlo). Note, however, that even this hypothetically perfect DA algorithm would not produce perfectly accurate and perfectly precise estimates of the true system state because of the fact that the total amount of information from the imperfect model and imperfect retrievals is not generally enough to fully specify the true state of the system.

More importantly, any realistic DA algorithm will not be perfectly efficient at extracting whatever amount of information does exist from the model and retrievals. Importantly, any assumptions or approximations in the DA algorithm itself (e.g., parametric approximations of uncertainty distributions and incomplete sampling) will result in some amount of information loss relative to the total that is hypothetically available from the model and retrievals.

Finally, when we test the results of any particular DA application against some set of evaluation data (usually in situ data), some portion of the mismatch between DA posteriors and evaluation data will be due to error in the evaluation data. As an example, ground truth soil moisture data might come from in situ probes that have effectively point-scale spatial support, whereas the model and remote sensing data might have spatial support on the order of tens or hundreds of meters. This means that we can never fully or precisely measure how well the DA results emulate the true system state, and this also means that we cannot fully measure the total information content of the model and retrievals about the true system state. Error in evaluation data might be systematic or random, and in particular, we generally expect that any mismatch between the spatiotemporal support of the evaluation data and the model grid will introduce systematic, site-specific biases, which must be accounted for when we quantify uncertainty in any DA system.

Taken together, these issues contribute to overall uncertainty in the results from any DA application; however, each of these issues requires a very different strategy for mitigation. On one hand, increasing the total amount of information available to a DA system requires either new model development to increase the realism or information content of the simulation model, or developing and deploying improved observing systems to increase the information content of the retrievals. On the other hand, improving a DA algorithm itself generally requires increasing computational expense (e.g., increasing ensemble size, eliminating parametric assumptions, implementing robust sampling methods like MCMC, dynamic tuning of model and retrieval error parameters or distributions). If we want to build or improve an existing DA system for a particular problem, then it would be beneficial to know the extent to which each of these three sources of uncertainty (model, retrievals, and algorithm) is a primary limiting factor on overall performance. Moreover, this three-way uncertainty segregation must be robust to the fact that there will generally be both random and systematic error in evaluation data; otherwise, we risk simply tuning a DA system to bad data.

In this paper, we propose a strategy for separating and quantifying the first three sources of DA uncertainty in the presence of the fourth. We refer to the ability for a given DA algorithm to use the full information content of models and retrievals as the *efficiency* of DA, and we will measure both total information content and information-use efficiency using information theory (Shannon, 1948). In particular, what we are calling efficiency is a form of statistical efficiency, which should not be confused with computational efficiency. In fact, our purpose is to propose a metric that could help quantify trade-offs between statistical and computational DA efficiencies in the presence of imperfect and incomplete information (i.e., uncertainty in models, retrievals, and evaluation data).

Our strategy is to first measure the information content of simulated model states, and of the retrieval data relative to the imperfect evaluation data, and then measure the fraction of this information that is extracted by a given DA implementation or algorithm. This allows us to measure empirically what fraction of total information loss could be mitigated in a particular DA application by spending more resources on improved observing systems versus improved computational DA algorithms. All of this is done without any assumption about how well imperfect evaluation data represents the true state of the system. The theory we propose is general, in the sense that it can be applied to any DA filtering application.

The rest of this paper is organized as follows. Section 2 outlines the theory, and section 2.1 in particular describes how to measure information contained in models and data, as well as the fraction of that information extracted by DA. Section 2.2 then describes how to use an information-theory divergence between Bayesian posteriors to segregate inefficiency effects due to different individual assumptions and/or approximations in the DA filter, so that we can identify specific causes of inefficiency in the DA algorithm. Section 3 gives an example application related to soil moisture, and section 4 contains a short discussion about what it means to say that a data fusion or inference strategy is *optimal*.

2. Theory

2.1. Information Use Efficiency

The situation that we investigate is one where a scientist intends to evaluate the performance of some particular DA application by comparison with a set of in situ measurements. These in situ measurements may be sparse or dense, have any spatiotemporal support, and may be directly or indirectly related to the variable that we are observing and assimilating. For example, we could conceivably evaluate assimilation of soil moisture retrievals by looking for improvements to modeled estimates of things like latent heat flux, leaf area index, future precipitation (via land/atmosphere couplings), or any other diagnostic variable.

Our problem therefore admits four primary variables: the random variable Y represents remote sensing retrievals (or whatever data is assimilated into the model), the random variable X represents model estimates of the state or flux that we want to estimate or predict without DA (this is called the *open-loop*), the random variable X^+ represents analysis estimates (after DA) of the same variable, and the random variable Z represents whatever measurements are used to evaluate the DA experiment. Again, variables X, Y, and Z do not necessarily represent precisely the same physical quantities, although X and X^+ do represent the same physical quantities before versus after DA.

Our notation will be that capital letters represent random variables and lowercase letters represent realizations of random variables. This holds for all variables except R and Q, which are standard notation for retrieval and model error covariances, respectively. In the remainder of this essay, the term *observation* refers exclusively to evaluation data, and *retrieval* refers to whatever data are assimilated into the model.

To address the main question outlined in section 1 about whether DA is limited by the information content of retrievals or by the ability of the assimilation algorithm to extract that information, we need to define a way to measure the information content of data. To begin, we propose to conceptualize the problem as illustrated in Figure 1. Loosely speaking, the area of each circle in these Venn diagrams represents our prior uncertainty, here measured as Shannon entropy, about one of our primary variables: X, Y, or Z. The area of the overlaps between the three circles represents the information shared by any pair or triplet of variables, here measured as mutual information. In general, information shared between a pair of variables is information contained in one variable that allows us to reduce uncertainty about the other variable.
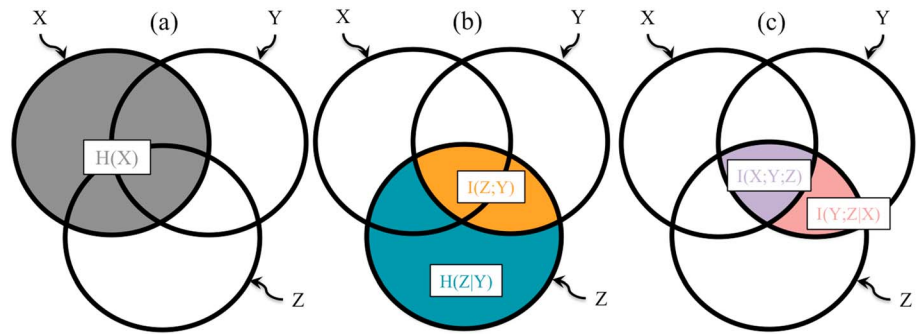
Figure 1. A representative illustration of the information metrics for a prototypical DA problem. The area occupied by any full circle (e.g., for X in (a)) represents the total entropy according to equation (2) of the corresponding variable: X = model predictions, Y = assimilated retrievals, and Z = in situ evaluation observations. The overlapping portions of the Venn diagrams (e.g., $I(X:Y)$ shaded in (b) represent the amount of information that is shared between each pair or triplet of variables according to equation (3). This shared information measures the amount of entropy that can be reduced in any one of the variables given knowledge of the other(s) according to equation (4.1).

2.1.1. Information-Theory Background

To make the diagrams in Figure 1 formal, we need quantities that represent the various components of these diagrams. This section (section 2.1.1) gives very brief background on the foundational information theory metrics, including mutual information and entropy, and more details can be found in the textbook by Cover and Thomas (1991). In the following subsections, we will use these standard information theory metrics to develop a new metric for DA efficiency.

Under probability theory, the change in our knowledge about one variable (e.g., Z) that occurs due to collecting and conditioning on new data (e.g., Y) is described by the ratio of the conditional distribution to the marginal distribution over the variable of interest:

$$\frac{p(Z|Y)}{p(Z)}. \tag{1}$$

Here $p(Z)$ is the probability distribution that represents everything we know about the evaluation observations Z before running a model or collecting remote sensing retrievals, and the conditional probability distribution $p(Z|Y)$ represents what we know about Z after considering only the retrievals. In our applications, $p(Z)$, $p(Y)$, and $p(Z, Y)$ are empirical, derived either as histograms or as kernel density functions.

We can integrate such ratios into metrics that represent the expected change in our knowledge about the value of any particular variable (e.g., Z) due to making direct measurements on that variable (i.e., measurements like $Z = z$). This statistic is called *entropy*, and is defined as

$$H(Z) = \int p(z) \ln\left(p(z)^{-1}\right) dz. \tag{2}$$

The entropy of the model predictions, $H(X)$, is illustrated by the gray-shaded region in Figure 1a. $H(X)$ is calculated by taking all of the open-loop model predictions during the entire assimilation period and domain and constructing an empirical distribution to get $p(X)$. Equation (2) is then applied to this empirical distribution. The process is similar for $H(Y)$ and $H(Z)$.

Suppose now that instead of having access to direct measurements like $Z = z$ we have access to measurements of a related variable—for example, some remote sensing retrievals $Y = y$. The next quantity we need is the expected divergence between what we would learn if we were to collect measurements like $Z = z$ depending on whether or not we had previously conditioned our knowledge of Z on information in measurements $Y = y$. This is quantified in a way that is analogous to equation (2) as the *mutual information* between Y and Z:

$$I(Y;Z) = \iint p(y,z) \ln\left(\frac{p(z|y)}{p(z)}\right) dy dz. \tag{3}$$

$I(Y;Z)$ can be interpreted as the expected amount of information about random variable Z that is contained in realizations of the random variable Y and vice versa; that is, $I(Y;Z) = I(Z;Y)$ always.

We now have everything necessary to formalize our diagrams in Figure 1. In particular, the expected residual entropy about the evaluation measurements Z conditional on retrievals Y is the difference between the two metrics above:

$$H(Z|Y) = H(Z) - I(Y;Z). \tag{4.1}$$

The mutual information statistic from equation (3) and the conditional entropy statistic from equation (4.1) are illustrated by the orange and blue-shaded regions in Figure 1b, respectively. Although we will not use this fact directly, it is worth noticing that since $I(Z;Y) = I(Y;Z)$, equation (4.1) is symmetric in the sense that

$$H(Y|Z) = H(Y) - I(Y;Z). \tag{4.2}$$

We expect that running a model is generally, or at least often, cheaper than deploying a new observing system, so if we want to know the expected marginal information content of the remote sensing retrievals over and above what is available from our model, then we are really interested in the probability ratio conditional on model predictions X:

$$\frac{p(Z|Y,X)}{p(Z|X)}. \tag{5}$$

To understand the information content of a remote sensing retrieval in the context of DA—where we are already running a model—we will need to estimate conditional information metrics like

$$I(Y;Z|X) = \iiint p(x,y,z) \ln\left(\frac{p(z|y,x)}{p(z|x)}\right) dx\,dy\,dz. \tag{6}$$

Equation (6) follows directly from equation (3) after three applications of the chain rule of probability theory (not shown here). $I(Y;Z|X)$ is the amount of information about Z gained by collecting retrievals in the case that we are already running a model to simulate X—this quantity is illustrated by the pink-shaded region in Figure 1c.

One might intuit that the amount of information contained in Y about Z conditional on X (i.e., $I(Y;Z|X)$) might be lower when conditional on a high-quality model than it would be if we were not running any model. The natural intuition is—we imagine—that some of the information from the retrievals will be redundant with some of the information from the model. This intuition is not necessarily true because of *information synergy* (Schneidman et al., 2003). It is possible for three-way mutual information statistics like $I(X;Y;Z)$, which represents the information shared by all three variables (purple-shaded region in Figure 1c), to be negative. Negative multivariate mutual information is somewhat unintuitive—it does not indicate that the variables are not informative of each other; rather, it means that any combination of two or more random variables together contain *more* information about a third random variable than the sum of the individual information contents of the predictor variables individually. That is, information synergy occurs when $I(Z;X,Y) = I(Z;X) + I(Z;Y) - I(X;Y;Z) > I(Z;X) + I(Z;Y)$, which only happens when $I(X;Y;Z) < 0$.

Our soil moisture DA examples in section 3 show evidence of synergistic information, and we expect that this will be relatively common in other DA applications as well. It is not necessarily the case that remote sensing retrievals become less valuable in a DA setting as the open-loop model accuracy improves.

2.1.2. Efficiency Metrics

We now have enough quantitative theory to derive an efficiency metric for any DA filter. The conceptual steps that we propose for this are illustrated in Figure 2. The information content of our open-loop predictions X about the evaluation data is given by $I(Z;X)$—this is the amount of entropy in the experimental variable that can be reduced by conditioning on model predictions; this concept is illustrated by the Venn diagram in Figure 2a. The retrievals Y presumably add some extra information on top of what is available from the model alone according to equation (6), and this is illustrated by the blue-shaded region in Figure 2b. We now have a total quantity of information available from the (imperfect) model and (imperfect) retrievals about the (imperfect) evaluation data that is given by the following sum rule:
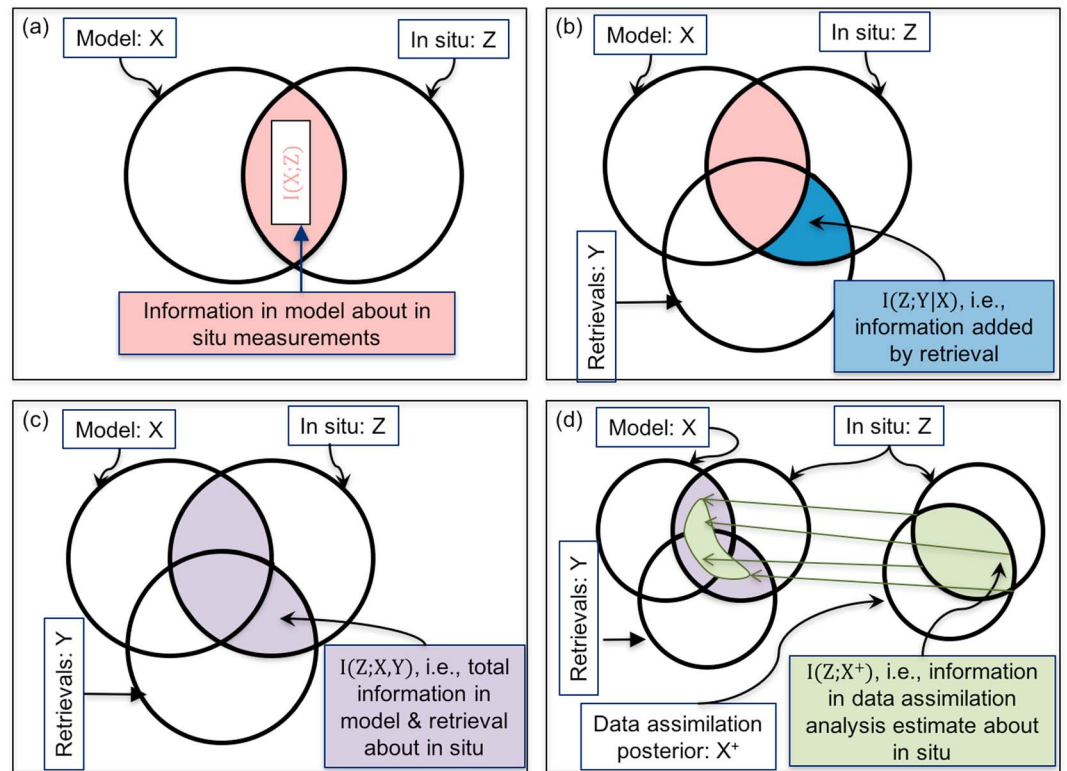
Figure 2. Conceptual diagram of a nonparametric DA efficiency metric. (a) The amount of information in the model about the in situ measurements according to equation (3) (with Y replaced by X). (b) The information added by the remote sensing retrievals according to equation (6). (c) The total amount of information captured by the model and retrieval together about the evaluation data according to equation (7). (d) The efficiency ratio in equation (8). The information shared between the DA analysis time series and the evaluation data will always be less than the total available information due to inefficiencies in the DA algorithm.

$$I(Z; X, Y) = I(Z; X) + I(Z; Y|X). \tag{7}$$

This quantity is illustrated by the purple-shaded region in Figure 2c, and is the total amount of information that is available to our DA filter in the context of our particular real-world experiment.

Next, we actually run the DA algorithm to obtain an analysis or posterior estimate of the model state, X^+. We can measure the information content about Z contained in the analysis estimate in a manner similar to equation (3) to obtain the quantity $I(Z; X^+)$. This quantity can never exceed the total available information from equation (7) due to the data processing inequality (Kinney & Atwal, 2014), and this is illustrated by the green-shaded region in Figure 2d. The data processing inequality states, simply, that if there exist a Markov conditioning relationship between three variables—say, for example, $X \rightarrow Y \rightarrow Z$—then $I(Z; X) \leq I(Z; Y)$ always. And important instance of this occurs when one random variable (or the probability distribution over that variable) is a function of another random variable (or the probability distribution over another variable). In our case, the Markov chain we care about is $X^+ \rightarrow \{X, Y\} \rightarrow Z$, which holds during DA because the probability distribution over X^+ is derived as a function of the joint probability distribution over $\{X, Y\}$; therefore, $I(Z; X^+) \leq I(Z; X, Y)$ always.

The first main assertion in this paper is that the ratio of the total information contained in the analysis to the total information available to the filter from the imperfect model and imperfect retrievals represents the information-use efficiency of DA:

$$\mathscr{E}_{DA} = \frac{I(Z; X^+)}{I(Z; X, Y)}. \tag{8}$$

Notice that if the model were able to perfectly simulate the evaluation data, then $I(Z;X) = H(Z)$ and $I(Z;X,Y) = H(Z)$ regardless of the quality of the remote sensing retrievals. In reality, the denominator of equation (8) will be less than the total entropy of the evaluation data due to model error and retrieval error. This efficiency concept is illustrated in Figure 2d.

Although equation (8) represents total DA information-use efficiency, we might instead be interested in the ability of DA to use information contained in retrievals specifically, rather than the ability of DA to use the total information available from the model plus retrievals. The information-use efficiency relative to retrievals only can be isolated by taking the information added by DA (as compared to using the model alone) as a ratio of the total information available from observations:

$$\mathcal{E}_Y = \frac{I(Z;X^+) - I(Z;X)}{I(Z;Y|X)}. \tag{9}$$

Equation (9) measures the increase of information in the analysis vector, as compared to the open-loop, due to DA as a fraction of the total information about evaluation data that are available from retrievals. The efficiencies in both equations (8) and (9) can be negative in the case that DA corrupts information in the models or retrievals (i.e., the analysis is worse than the open-loop when compared against the evaluation data), but they can never exceed a value of 1.

2.1.3. Properties of the Efficiency Metrics

To reiterate, these efficiency metrics do not assume that the X, Y, and Z variables in equations (8) and (9) all represent the same physical quantities. This includes whether or not the model, retrieval, and evaluation data all have similar spatiotemporal support or spatiotemporal resolution. For example, it is a common problem when evaluating remote sensing data that in situ evaluation data do not necessarily have the same spatial support as the retrieval. As an example, X might be soil moisture simulated by a model at a finite resolution, and Z might be in situ soil moisture data, effectively at a point scale. Or X and Z could represent different physical variables altogether; perhaps X is soil moisture simulated on a 1/8° grid (e.g., Xia et al., 2013), and Z is leaf area index as estimated from remote sensing (e.g., Knyazikhin et al., 1999). Nevertheless, there is some amount of information shared between the modeled data, the retrievals, and the in situ data, and this total amount of information can, in principle, be extracted by a hypothetically *perfect* DA algorithm in the analysis vector X^+. Our efficiency metrics measure only the ability of DA to extract whatever information is available in the imperfect data sources.

To formalize this, consider three Markov chains such that there is some set of underlying *true* physical variable θ that exerts causal influence on three other true physical variables, say ξ, ψ, and ζ, which are each either observed or modeled by X, Y, and Z, respectively. ξ, ψ, and ζ could be the same variable (e.g., soil moisture) at different spatiotemporal scales, or could be different variables altogether (e.g., LAI and soil moisture). The resulting Markov chains can be expressed as $\theta \rightarrow \xi \rightarrow X$, $\theta \rightarrow \psi \rightarrow Y$, and $\theta \rightarrow \zeta \rightarrow Z$.

Ideally, we would want our efficiency metrics E_{DA} and E_Y, which are calculated with imperfect evaluation data Z, to be either bounded or convergent with hypothetical versions of these same efficiency metrics that would result if we were able to measure exactly the underlying true state of the system (i.e., that would result if $Z = \theta$ exactly). Bounding these metrics does not appear to be possible, since this would depend on there being a consistent ordering between the ratios $\frac{I(Z;X^+;\theta)}{I(Z;X,Y;\theta)}$ and $\frac{I(X^+;\theta)}{I(X,Y;\theta)}$, which does not follow from any reasonable set of assumptions.

We cannot have boundedness; however, the assumption that gives our efficiency metrics an intuitive interpretation is that each variable is related to the others only through the true state of the system—that is, there are no spurious, persistent correlations between X, Y, and Z. This assumption about spurious correlations means that the only correlation between X and Z (Y and Z) is through θ (the true state of the system), which we formalize by saying that (a) $p(X|Z,\zeta) = p(X|\zeta)$ and (b) $p(Z|X,\xi) = p(Z|\xi)$, and similarly for Y. It is always true that $I(Z;X,\xi) \geq I(Z;X)$, and the consequence of (b) is that $I(Z;X,\xi) = I(Z;\xi)$; taken together, these imply that $I(Z;\xi) \geq I(Z;X)$. This means that as uncertainty in the simulation/observation of ξ (ψ) by X (Y) reduces, these simulations (observations) cannot provide any more information about Z than would be available if we knew the true physical variables (i.e., ξ and ψ) exactly.

Formally, the efficiency metrics have two desirable properties:

1. Information in the DA analysis about the evaluation data is bounded above by total information in the model and retrievals: $I(Z; X^+) \leq I(Z; X, Y)$. This means that $\mathscr{E}_{DA}$ and $\mathscr{E}_Y$ are bounded above by 1.
2. Perfect measured efficiency relative to (imperfect) evaluation data (i.e., $\mathscr{E}_{DA} = 1$ or $\mathscr{E}_Y = 1$) only occur when perfect information extraction is achieved by the data assimilation algorithm relative to the (unknown) true state of the system.

Notice that this does *not* mean that perfect measured efficiency relative to (imperfect) evaluation data (i.e., $E_{DA} = 1$ or $\mathscr{E}_Y = 1$) only occurs when perfect information extraction is achieved with respect to the true state of the system.

As mentioned in section 2.1.2, the first condition is always true because of the data processing inequality (Kinney & Atwal, 2014). Because X^+ depends completely on X and Y and not at all on Z (i.e., the DA algorithm does not ingest the evaluation data directly), these variables have a Markov relationship like $Z \rightarrow \{X, Y\} \rightarrow X^+$. By the data processing inequality, we therefore know that $I(Z; X^+) \leq I(Z; X, Y)$ and $E \leq 1$. Similarly, since $I(Z; X, Y) = I(Z; Y|X) + I(Z; X)$, then $I(Z; X^+) - I(Z; X) \leq I(Z; Y|X)$ and $\mathscr{E}_Y \leq 1$. Thus, the first condition is always satisfied as long as the evaluation data, Z, are not used in the DA procedure.

The second condition is formally expressed as $I(Z; X^+) = I(Z; X, Y) \rightarrow I(\theta; X^+) = I(\theta; X, Y)$, where the arrow here is the symbol for logical implication, rather than the symbol for a probabilistic conditioning relationship as in the various Markov chain expressions. This condition is met under the assumption that any errors in the (imperfect) evaluation data relative to the (unknown) true state of the system are independent of any errors in the (imperfect) model and (imperfect) retrieval, conditional on the (unknown) truth. Stated formally, we assume that the following Markov chains hold $X \rightarrow \theta \rightarrow Z$ and $Y \rightarrow \theta \rightarrow Z$, so that $\{X, Y\} \rightarrow \theta \rightarrow Z$. Because X^+ is a mapping from X and Y and is not directly dependent on Z, our four random variables have the following Markov relationship: $X^+ \rightarrow \{X, Y\} \rightarrow \theta \rightarrow Z$. We want to show that the condition of measuring perfect efficiency, that is, $I(Z; X^+) = I(Z; X, Y)$, implies bona fide perfect efficiency, that is, that $I(\theta; X^+) = I(\theta; X, Y)$. This is easy to see, since the condition $I(Z; X^+) = I(Z; X, Y)$ implies that the following Markov property also holds: $\{X, Y\} \rightarrow X^+ \rightarrow \theta \rightarrow Z$, and thus, by the data processing inequality, it is necessarily the case that $I(\theta; X^+) = I(\theta; X, Y)$.

To summarize, our efficiency metrics in equations (8) and (9) have the natural property that they are bounded above by $E_{DA} = 1$ and $\mathscr{E}_Y = 1$ and that they only achieve this bound when the DA algorithm is perfect at extracting information *about the unobserved (and typically unobservable) true state of the system*. This is true no matter the quality of our evaluation data Z, and no matter the relationship between our evaluation data and the model simulations or retrievals—even if the evaluation data is at a completely different scale or of a completely different variable than the model simulations or retrieval data.

2.1.4. Empirical Estimators

All of the above entropy and information metrics can be estimated from samples of (i) the three-way joint distribution between X, Y, and Z and (ii) samples of the two-way joint distribution between X^+ and Z. This is done by using empirical distributions like $p(X, Y, Z)$ and Monte Carlo integration of equations like equations (2), (3), and (6). Paninski (2003) gives a review of some simple empirical estimators for entropy and mutual information metrics, and we use those estimators here.

One advantage of using discrete-entropy metrics (e.g., Nearing et al., 2013) is that this results in entropy and mutual information measures that are bounded below by zero, and thus, equations (8) and (9) have straightforward interpretations as standard efficiency metrics (bounded above by 1). In this case, the integrals in all of the entropy and mutual information equations are sums over discretizations of the random variable, and it is important that we are careful about how to discretize the random variables. Most importantly, it is necessary that our discretization is not too fine relative to the volume of available data, so that the empirical probability distributions are not degenerate. In our application example in section 3, we will analyze the precision of these metrics estimated from a finite data record.

2.2. Decomposition of Bayesian Data Assimilation

We have, in equations (8) and (9), measures of information-use efficiency for a DA filter. In cases where equation (8) returns a value of $E_{DA} < 1$, we would like to know what is causing the filter to be unable to extract all of the information contained in our assimilated retrievals.

To accomplish this, we will use the fact that all DA methods are fundamentally approximations of Bayes' theorem (van Leeuwen, 2010; Wikle & Berliner, 2007). The most basic statement of a DA filter is

$$p_a(X_t|y_t, x_{1:t-1}) = \frac{p(y_t|X_t)p(X_t|x_{1:t-1})}{\int p(y_t|X_t)p(X_t|x_{1:t-1})dX_t}. \tag{10}$$

It is important to understand that the posterior distribution p_a (called the analysis distribution) is our best estimate of the true value of the system component simulated by X (i.e., ξ from section 2.1.3) given information from the model and the retrieval.

There is a natural similarity between the filter expression in equation (10) and the mutual information metrics discussed in section 2.1. In this section, we will exploit that similarity to develop some diagnostic tests for the *causes* of information loss in an imperfect DA filter.

To illustrate this, let us start by using a more precise notation of the DA filter equation:

$$p_a(\xi_t|y_t, x_{1:t-1}) = \frac{p(y_t|\xi_t)\int p(\xi_t|X_t)p(X_t|x_{1:t-1})dX_t}{\int p(y_t|\xi_t)\int p(\xi_t|X_t)p(X_t|x_{1:t-1})dX_t d\xi_t}. \tag{11.1}$$

Notice that the likelihood $p(y_t|\xi_t)$ in equation (11.1) recognizes independence between the retrieval and the model conditional on the true state of the system ξ, which was our primary assumption from section 2.1.3. Equations (10) and (11.1) are similar only if we make the assumption that the modeled state of the system X_t is a random variable representing our best estimate (conditional on X, and Y) of the true state of the system ξ at time t, but this is exactly the assumption we are testing when using the efficiency metrics outlined in equations (8) and (9). So we must understand a DA filter using equation (11.1) rather than equation (10).

Further, notice that if we use a deterministic state-updating model m such that $x_t = m(x_{t-1}, \ldots)$, and the distribution $p(X_t|x_{1:t-1})$ is such that $X_t \mid x_{1:t-1} \sim f(x_t, \ldots)$, then we can re-write equation (11.1) as

$$p_a(\xi_t|y_t, x_{1:t-1}) = \frac{p(y_t|\xi_t)p(\xi_t|x_t = m(x_{t-1}, \ldots))}{\int p(y_t|\xi_t)p(\xi_t|x_t)d\xi_t}. \tag{11.2}$$

The key insight is that the conditional mutual information metrics from section 2.1 are derived from a similar application of Bayes' theorem:

$$p(Z_t|y_t, x_t) = \frac{p(y_t|Z_t, x_t)p(Z_t|x_t)}{\int p(y_t|Z_t, x_t)p(Z_t|x_t)dZ_t}. \tag{12}$$

In this case, we cannot assume conditional independence in the likelihood term like we did in equations (11.1) and (11.2) because, in general, we expect that the model and retrieval will share information that is not contained in the evaluation observations; that is, $I(X; Y|Z) > 0$. The conditional distribution on the left-hand side of equation (12) is used in equation (6) to calculate the information contained in the retrievals conditional on the model $I(Y; Z|X)$.

We now have the basic tools we need to diagnose causes of inefficiencies in any particular application of any DA filter evaluated against any particular set of in situ data. We will explain the procedure by example. In section 1 we said that it is common to use the EnKF for soil moisture data assimilation. The EnKF contains several distinct assumptions about linearity and Gaussianity of various joint and conditional distributions. In particular, the EnKF is *optimal* only in the case where (i) retrievals are linearly related to the modeled state, (ii) retrieval errors are Gaussian distributed, and (iii) errors in all modeled states are jointly Gaussian distributed. In cases where these assumptions do not hold, the EnKF will be inefficient, in the sense of equations (8) and (9), at extracting information from retrievals.

Suppose that we want to test the effects on a particular DA experiment due to assuming that the likelihood function (i.e., the observation operator) is linear-Gaussian when this relationship really is not linear or the uncertainty really is not Gaussian. In this case, we would substitute into the right-hand side of equation (12) the assumed likelihood function, which in this case is Gaussian with a prescribed retrieval error covariance R:

$$p_{\mathcal{N}}(Z_t|y_t, x_t) = \frac{\mathcal{N}(y_t|Z_t, R)p(Z_t|x_t)}{\int \mathcal{N}(y_t|Z_t, R)p(Z_t|x_t)dZ_t}. \tag{13}$$

$N(.|\mu, \Sigma)$ notates a Gaussian with mean μ and covariance Σ. In this case, $p_{\mathcal{N}}$ is not the analysis distribution from the EnKF, but instead is the posterior from equation (12), but with some of the standard EnKF assumptions built in. We now measure the expected information loss due to the normality assumptions in the likelihood as the divergence *from* the data-derived conditional *to* the (partially) analytic conditional from equation (13):

$$D(p\|p_{\mathcal{N}}) = \iiint p(x, y, z) \ln\left(\frac{p(z|y, x)}{p_{\mathcal{N}}(z|y, x)}\right) dxdydz. \tag{14}$$

Notice that this type of divergence is directly analogous to the mutual information metrics used previously—the mutual information metric in equation (3) is just a divergence from a marginal distribution to a conditional distribution over the same random variable. Here in equation (14), we are measuring a divergence between two different conditional distributions over the same random variable. Notice also that if we were to use a DA filter other than the EnKF, the specific assumption about the form of the likelihood function would be different, and so an analogous version of equation (14) would result.

Equation (14) will return zero divergence when the empirical distribution $p(y_t|Z_t, x_t)$ is (i) exactly Gaussian with covariance R and (ii) when Y is independent of X conditional on Z; that is, when $p(y_t|Z_t, x_t) = \mathcal{N}(y_t|Z_t, R)$. We can, of course, test the second assumption independent of the first by measuring the appropriate information loss as

$$D(p\|p_I) = \iiint p(x, y, z) \ln\left(\frac{p(z|y, x)}{p_I(z|y, x)}\right) dxdydz \tag{15.1}$$

$$p_I(Z_t|y_t, x_t) = \frac{p(y_t|Z_t)p(Z_t|x_t)}{\int p(y_t|Z_t)p(Z_t|x_t)dZ_t}. \tag{15.2}$$

The only difference between equations (15.2) and (12) is that we replaced $p(y_t|Z_t, x_t)$ with $p(y_t|Z_t)$, and p_I notates the distribution that results from an *independence* assumption. Like $p(y_t|Z_t, x_t)$ in equation (12), $p(y_t|Z_t)$ is derived directly from data. Again, if we were testing a different DA filter, we would use a different substitution for the empirical likelihood function.

The key takeaway is that $D(p\|p_{\mathcal{N}}) \geq 0$ measures the extent to which real-world violations of the EnKF normality and/or conditional independence assumptions in the observation operator (i.e., the likelihood function) contribute to information loss in this particular application experiment, while $D(p\|p_I) \geq 0$ measures only the information loss due to the conditional independence assumption in the likelihood function. If it were exactly true that (1) our retrievals were Gaussian distributed around our evaluation measurements with covariance R and (2) that Y is independent of X conditional on Z, then equations (14) and/or (15.1) would return $D(p\|p_{\mathcal{N}}) = 0$ and $D(p\|p_I) \geq 0$ indicating zero information loss.

Given that we are interested in separating inefficiency effects due to artifacts in the filter prior and likelihood, it is useful to notice that the divergences of the decomposed joint distributions are additive. To state this generally, imagine applications of Bayes' theorem to two different joint probability distributions over two random variables Y and Z, $p(Y, Z)$ and $q(Y, Z)$; in this general case we have the following result:

$$D(p(Y|Z)p(Z)\|q(Y|Z)q(Z)) = D(p(Y|Z)\|q(Y|Z)) + D(p(Z)\|q(Z)). \tag{16}$$

In the context of DA, this means that if we use the EnKF prior and likelihood, then the information loss by the EnKF (equation (10)) relative to the data-derived posterior (equation (13)) is decomposed as

$$D(p(Z_t|y_t, x_t)\|p_a(Z_t|y_t, x_{1:t-1})) = k + D(p(y_t|Z_t, x_t)\|\mathcal{N}(y_t|Z_t, R)) + D\left(p(Z_t|x_t)\|\mathcal{N}\left(Z_t|\overline{X_t}, \overline{Q_t}\right)\right). \tag{17}$$

$\overline{X_t}$ and $\overline{Q_t}$ are the ensemble mean and covariance at time t, and k is a constant related to the Bayesian normalizing factors. Notice that the $\mathcal{N}()$ terms on the right-hand side of the equation arise from the EnKF normality assumptions, and different assumptions would result if a different DA filter were used.
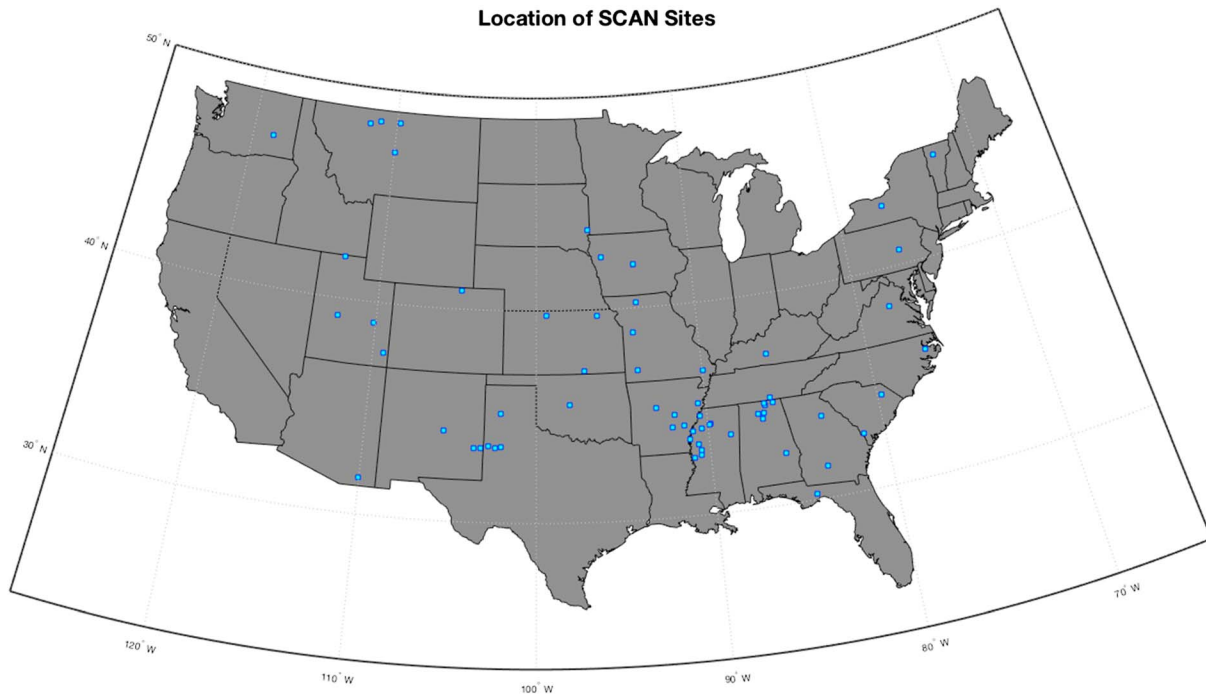
Figure 3. The 64 SCAN site locations.

We will not directly use equations (16) and (17), but they are useful to develop an intuition about how these divergence decompositions relate to the efficiency metrics in equations (8) and (9). Notice that the information loss from equation (8) is equal to the total divergence in equation (17):

$$I(Z; X, Y) - I(Z; X^+) = D(p(Z_t|y_t, x_t) \| p_a(Z_t|y_t, x_{1:t-1})) = \iiint p(x, y, z) \ln\left(\frac{p(z|y, x)}{p(z|x^+(x, y))}\right) dx\,dy\,dz. \tag{18}$$

We can use this basic strategy to measure information loss—up to a constant k related to the Bayes normalizing constants—due to *any individual assumption or approximation* in the DA filter. The individual parametric or probabilistic assumptions in the likelihood and/or prior will depend on the particular DA algorithm we want to test.

3. Example Application: Methods and Results

This section presents an experiment that uses the theory outlined in section 2 to quantify the efficiency of particular applications of the EnKF to assimilate soil moisture retrievals. The objective of this example is to mimic, as closely as possible, the experimental setup of an existing soil moisture DA experiment—in particular, we chose to mimic the experimental setup by Kumar et al. (2014). Our experiment first measures the efficiency of Kumar et al.'s DA strategy using the efficiency metrics described in section 2.1, and then diagnoses the causes of inefficiency using the divergence metrics outlined in section 2.2. Finally, we ran a set of experiments that look at the effect on information loss of changing the parameters of the EnKF, specifically the assumed retrieval error covariance, and the state perturbation covariance.

3.1. Models, Data, and Assimilation Details

Soil moisture retrievals from the Land Parameter Retrieval Model (LPRM; Owe et al., 2008), which is based on observations from the Advanced Microwave Scattering Radiometer for Earth (AMSR-E), were assimilated into the Noah Multi-Parameterization (Noah-MP) land surface model (Niu et al., 2011) for the time period of 2001–2011. Results were evaluated against point-based in situ data from 64 of the USDA Soil Climate Analysis Network (SCAN; Schaefer et al., 2007) sites; these are illustrated in Figure 3. Evaluation data from SCAN were made at a depth of 2 inches (5 cm).

Table 1
Noah-MP Configuration Options

Physical process	Available Noah-MP options	Option used here
Vegetation	• Prescribed LAI and shade fraction • LAI and shade fraction from dynamic carbon uptake and partitioning • Shade fraction calculated from prescribed LAI • Prescribed LAI and constant shade fraction	Prescribed LAI and shade fraction
Stomatal resistance	• Ball-Berry (Ball et al., 1987) • Jarvis (Chen et al., 1996)	Ball-Berry
Soil moisture factor for stomatal resistance	• Noah-type (based on soil moisture) • CLM-type (based on stomatal resistance) (Oleson et al., 2010) • SSiB-type (based on stomatal resistance) (Xue et al., 1991)	Noah-type
Runoff and groundwater	• TOPMODEL with groundwater (Niu et al., 2007) • TOPMODEL with equilibrium water table (Niu et al., 2005) • Infiltration-excess surface runoff and free drainage (Schaake et al., 1996) • BATS runoff and free drainage (Yang & Dickinson, 1996)	TOPMODEL with groundwater
Surface layer drag coefficient	• Monin-Obukhov • Noah-type (Chen et al., 1997)	Monin-Obukhov
Super-cooled liquid water	• No iteration (Niu & Yang, 2006) • With iteration (Koren et al., 1999)	No iteration
Frozen soil permeability	• Linear: hydraulic properties from total soil moisture (Niu & Yang, 2006) • Nonlinear: hydraulic properties from liquid water only (Koren et al., 1999)	Linear: total soil moisture
Radiation transfer	• Two-stream with 3-D structure • Two-stream (Niu & Yang, 2004) • Two-stream with canopy gap equal to 1-(shade fraction)	Two-stream with 3-D structure
Snow albedo	• BATS (snow age, grain size growth, impurity) (Yang et al., 1997) • CLASS (only snow age) (Verseghy, 1991)	CLASS
Frozen/liquid partitioning	• Jordan (1991) • Offset threshold: $T_{air} < T_{frz} + 2.2$ K • Standard threshold: $T_{air} < T_{frz}$	Jordan (1991)
Bottom soil temperature	• Zero heat flux • Prescribed (8 m) bottom temperature	Prescribed (8 m) bottom temperature
Soil temperature solution	• Semi-implicit • Full implicit	Semi-implicit

Noah-MP was run at an hourly time step over four soil layers with depths of 5, 10, 35, and 150 cm. Model parameters and forcing data were from NLDAS (Xia et al., 2013). Noah-MP configuration options are listed in Table 1. These configuration options are explained in some detail by Niu et al. (2011).

The EnKF was used to update the volumetric soil water content in the four modeled soil layers given LPRM retrievals. LPRM retrievals are given in units of volumetric water (m³/m³) with approximate retrieval depth of 5 cm, and with spatial resolution of approximately a quarter degree. Cumulative density function (CDF) matching is necessary to remove biases between the model and retrievals (Kumar et al., 2012), and will—at least potentially—result in some amount of information loss, which we will measure directly. We applied CDF matching to transform retrievals into the Noah-MP model climatology using a discrete binned transform with resolution of 1% of the range of the modeled and observed values. CDF matching was implemented independently at each of the 64 sites (but did not use SCAN data).

The EnKF observation operator was an identity operator on the top-layer soil moisture with retrieval error covariances supplied by standard LPRM data files (all retrieval error covariances were either $\sqrt{R} = 0.06$ m³/m³ or $\sqrt{R} = 0.10$ m³/m³). Retrieval error covariances were scaled at each site based on the changes to the total variance of the retrievals due to CDF matching: $R^* = R \frac{\text{var}(Y)}{\text{var}(Y_{cdf})}$, where Y_{cdf} are the transformed observations, but we allowed no scaled retrieval error covariance to be less than $\sqrt{R} = 0.01$ m³/m³. This means that the retrieval errors were different at each site. The EnKF ensemble consisted of 50 members. Ensemble forcing perturbations were sampled from the cross-correlated distribution used by Kumar et al. (2014), which are listed in Table 2.

Table 2
EnKF Forcing and State Perturbations (50 Ensemble Members)

Variable	Perturbation type	Standard deviation	Cross correlations			
Forcings			SW	LW		Precipitation
Shortwave radiation	Multiplicative	0.3	1.0	−0.5		−0.8
Longwave radiation	Additive	50 W/m^2	−0.5	1.0		0.5
Precipitation	Multiplicative	0.5	−0.8	0.5		1.0
Soil moisture states			Layer 1	Layer 2	Layer 3	Layer 4
Layer 1 (5 cm)	Additive	6×10^{-3} m^3/m^3	1.0	0.6	0.4	0.2
Layer 2 (10 cm)	Additive	1.1×10^{-4} m^3/m^3	0.6	1.0	0.6	0.4
Layer 3 (35 cm)	Additive	6×10^{-6} m^3/m^3	0.4	0.6	1.0	0.6
Layer 4 (150 cm)	Additive	4×10^{-6} m^3/m^3	0.2	0.4	0.6	1.0

### 3.2. EnKF Efficiency Experiments

The first step in the efficiency analysis is to collect a single set of coincident (collocated and co-temporal) open-loop model predictions, satellite retrievals, in situ measurements, and EnKF analysis predictions: $\left\{ X_{i,t}, Y_{i,t}, Z_{i,t}, X_{i,t}^+ \right\}_{i=1,\dots,m;t=1,\dots,t_i}$, where $m = 64$ is the number of in situ measurement locations—illustrated in Figure 3. The 1200 sample points were sampled from the complete 2001–2011 hourly time series of each of the $m = 64$ SCAN locations, for a total of $N = 76,800$ sample points. All entropy and information statistics were estimated from these N samples.

It is important to reiterate that it is necessary to include all sites together in the calculation of a single set of information metrics. If we used a site-by-site or site-independent analysis—for example, calculating equations (3), (6), (8), and/or (9) individually for each SCAN site, then the resulting statistics would be influenced by any systematic bias between the true surface soil moisture within each pixel and the soil moisture measured at the specific in situ location within that pixel. For example, we would not, in that case, measure the information content of the retrievals that could help mitigate additive bias in the soil moisture state of the model. Additionally, if we treated each site independently, then we would run the risk of overestimating the information content of the models and/or retrievals since the data at an individual site may imply systematic relationships between the point-based in situ measurement and the retrieval or model. We cannot know such site-specific systematic errors, and therefore must calculate information statistics by integrating over all of the available sites. This spatial integration is implicit in the integrations over empirical probability distributions in equations (3) and (6) as long as we include data pairs and triplets from all sites when we estimate the relevant empirical joint distributions.

Before we present the results of our efficiency experiments, we must briefly explore the implications of sample size and data resolution. All empirical probability distributions were derived as histograms with fixed bin widths, and maximum likelihood estimators of nonparametric discrete entropy and mutual information statistics like these are convergent but biased (Paninski, 2003). We must choose a histogram resolution that allows for robust statistical estimators given our available sample size. Figure 4 shows the effect of sample size and bin resolution on the EnKF efficiencies from equation (9), and the smallest discretization that allows for a stable efficiency estimator at $N = 76,800$ is about 0.03-m^3/m^3 volumetric soil moisture.

Table 3 presents the primary results of these efficiency experiments. The statistics in this table were calculated from the full data sample using a discretization of 0.03 m^3/m^3, which was chosen according to Figure 4. The first thing to notice is that the total available information from model and retrievals together generally explains less than 20% of the total entropy of the in situ data (fourth row in Table 3). The second thing to notice is that the retrievals alone generally contain marginal information of about 5% of the total variability of the SCAN data (third row in Table 3). This means that, a priori, we will expect generally relatively small improvements to our ability to inform SCAN data after assimilating LPRM retrievals.

To put this in some perspective, soil moisture retrievals are generally evaluated against in situ networks using linear or second-order statistics like the product–moment correlation coefficient or a fraction of explained variance (e.g., Liu et al., 2011). Theoretical relationships between second-order statistics and information theory metrics were derived numerically for the ideal second-order case (i.e., all probability distributions are
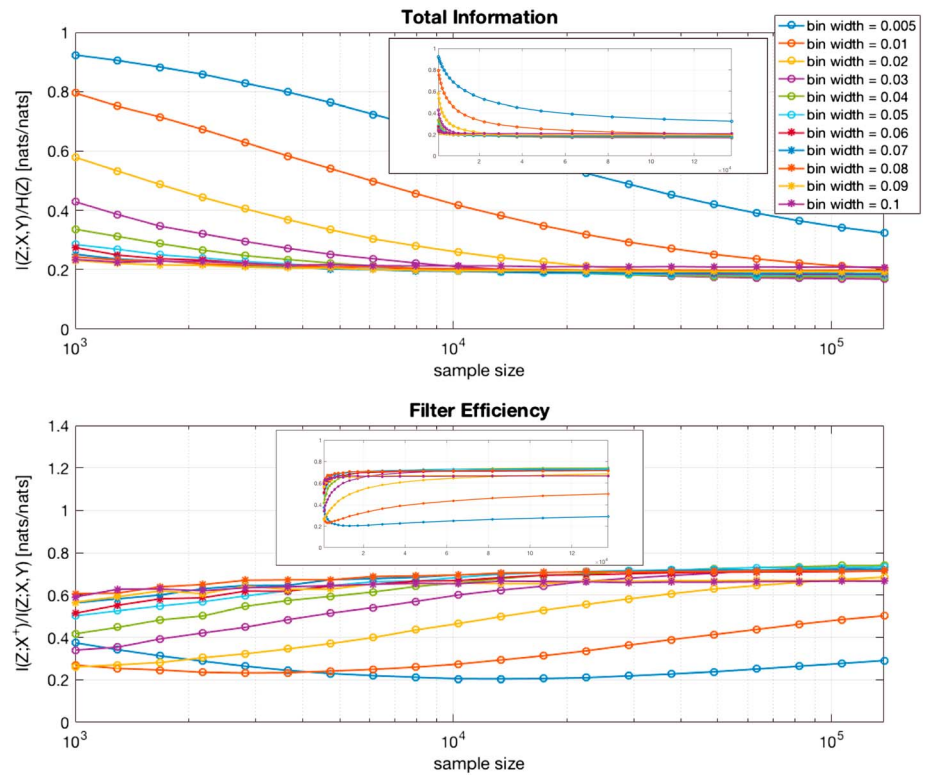
Figure 4. Convergence of total information, $I(Z; X, Y)$, and EnKF efficiency, $\mathcal{E}_{DA}$, as a function of sample size at different histogram bin widths used for estimating the empirical joint distribution $p(X, Y, Z)$. The inserts show the same data as the primary plots, but plotted on a linear x axis.

Table 3
Breakdown of the Information-Use Efficiency Metrics From an EnKF Assimilation of LPRM Retrievals Into the Noah-MP Land Surface Model as Evaluated Against SCAN Data

Measurement	Metric	Value
Information in model simulations[a]	$\dfrac{I(Z;X)}{H(Z)}$	0.13
Information in retrievals[a]	$\dfrac{I(Z;Y)}{H(Z)}$	0.08
Conditional information in retrievals[a]	$\dfrac{I(Z;Y|X)}{H(Z)}$	0.05
Total information from model and retrievals[a]	$\dfrac{I(Z;X,Y)}{H(Z)}$	0.18
Fraction of retrieval information lost via CDF-matching	$1 - \dfrac{I\left(Z;Y^{CDF}|X\right)}{I(Z;Y|X)}$	0.11
Information from EnKF	$\dfrac{I(Z;X^{+})}{H(Z)}$	0.13
Efficiency of EnKF $\left(\mathcal{E}_{DA}\right)$	$\dfrac{I(Z;X^{+})}{I(Z;X,Y)}$	0.72
Efficiency of EnKF $\left(\mathcal{E}_{y}\right)$	$\dfrac{I(Z;X^{+})-I(Z;X)}{I(Z;Y|X)}$	0.03

[a]Information metrics are normalized by the total entropy of the evaluation data Z so that their values range between 0 and 1. They are interpreted as, for example, "the fraction of total uncertainty about measurements Z that can be resolved given Y and X."
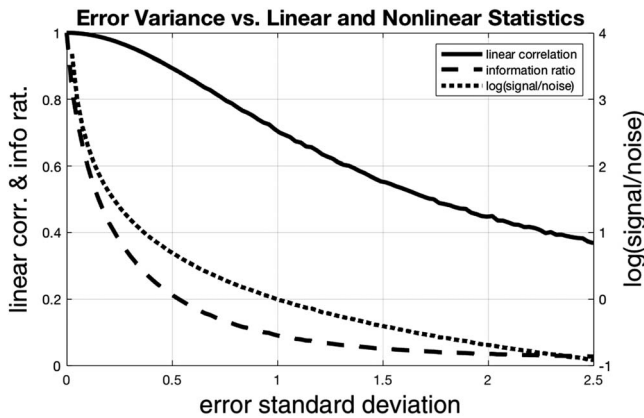
Figure 5. The general relationship between standard linear-Gaussian statistics and our fraction of explained information statistics like $I(Z; X)/H(Z)$. These theoretical relationships are valid for standard normal predictands and predictors with additive Gaussian noise, and the signal-to-noise ratio is defined as the ratio of the root variance of the predicted variable to the root variance of the error in the predictor (i.e., a signal-to-noise ratio of 1 means that the variance of the error in the predictor is equal to the variance of the variable being predicted).

jointly Gaussian)—these are shown in Figure 5. Notice that information ratios follow much more closely the variance-based signal-to-noise ratio than does the product–moment correlation coefficient. As a point of reference, a linear correlation coefficient of $\rho^2 = 0.71$ corresponds to a signal-to-noise ratio of 1—these both happen at an error standard deviation of 0.5.

In our example here, with a bin resolution in the empirical joint distributions of 0.03 m^3/m^3, an information statistic with value of $\frac{I}{H} = 0.05$ represents a linear correlation of $\rho^2 \approx 0.55$, which occurs in a linear model when the error standard deviation is approximately 1.5 times the signal standard deviation. This is higher than the $\rho^2 = 0.42$ found by Liu et al. (2011) when comparing AMSR-E retrievals with a subset of data from the SCAN network, and similar to the $\rho^2 = 0.55$ that they found when comparing AMSR-E retrievals against cal/val core site data. Our retrieval statistics generally agree with those published previously, and the point is that if we are going to use point-scale data to evaluate DA applications (e.g., Kumar et al., 2014), we have to be aware that it appears to be typical that there is very little information in retrievals about point-scale phenomena to begin with.

Liu et al. (2011) found that the fraction of variance in SCAN in situ measurements that was explainable by AMSR-E retrievals was essentially equal to the fraction of variance explainable by the Catchment land model forced with MERRA data ($\rho^2 = 0.42$ versus $\rho^2 = 0.43$). This disagrees somewhat with our analysis, which shows that Noah-MP provides more information than the AMSR-E retrievals. Information fractions of $\frac{I}{H} = 0.13$ for the model versus $\frac{I}{H} = 0.08$ for the retrievals would correspond to linear correlation coefficients of $\rho^2 \approx 0.8$ versus $\rho^2 \approx 0.67$ if the relationships were actually linear-Gaussian (again, see Figure 5). This difference could be due to either the fact that (1) our model (Noah-MP) and forcing data (NLDAS) are better than the Catchment model with MERRA forcing data used by Liu et al. or that (2) our nonparametric information metrics capture some nonlinear portion of the signal in the model predictions that is not present in the retrievals. In fact, our Noah-MP correlation coefficient is $\rho^2 \approx 0.68$ and our LRPM correlation coefficient is $\rho^2 \approx 0.54$—our Noah-MP model forced by NLDAS is somewhat better than Liu et al.'s Catchment/MERRA configuration, but the nonparametric measure is capturing some systematic nonlinear relationship between Noah-MP and SCAN soil moisture.

The next important result is that CDF-matching cost about 10% of the conditional information in the retrievals (fifth row in Table 3). It is interesting to notice that histogram matching is an invertible transform, and should therefore be information preserving. However, CDF matching was applied locally, and we calculated our information statistics globally across all sites in Figure 3. Information loss due to CDF matching is due to the fact that these local transforms all have different inversions, but all of the data are lumped in the same empirical density functions for the integrals in equations (3) and (6). There is no single mapping to invert the localized CDF matching that works across all sites.

The most important results for our purposes are about the efficiency of the EnKF (lines 7 and 8 of Table 3). The efficiency of the EnKF in this particular application at extracting information from retrievals is less than 5% (line 8 of Table 3). It is certainly true that there is not a lot of information in the retrievals to begin with; however, of that small amount of information, the EnKF uses only 5%. Almost no information is extracted from LPRM observations in this example.

Notice that the value of the total efficiency metric in Table 3 ($E_{DA} = 0.72$) is relatively high. This means that the EnKF in this case was able to extract more than three quarters of the total information available from the Noah model and LPRM retrievals. This might seem high, except that the model itself provides $\frac{0.13}{0.18} = 72\%$ of the total available information. So a value of $\mathscr{E}_{DA} = 0.72$ simply means that the EnKF did not lose information relative to the open-loop (no assimilation). Similarly, Kumar et al. found very little improvement over the open-loop using their squared-error and correlation statistics, and what our metrics add to their analysis is to address whether the poor improvement is due to high uncertainty (low information) in the LPRM data versus

to information loss in the DA algorithm. The answer to this question comes from the $\mathscr{E}_Y = 0.03$ efficiency metric. These results show that while it is certainly true that there is very little marginal information in the LPRM retrievals (over and above what is available from the Noah model), it is also the case that the DA algorithm is able to extract only a small amount of what little information exists.

3.3. EnKF Decomposition Experiments

We have seen that our EnKF implementations are inefficient in this particular example—meaning that they did not make full use of the available information from the remote sensing retrievals. Now let us try to understand what exactly is causing this information loss.

To begin, we calculated the divergence (from equation (14)) from the empirical conditional $p(Z_t|y_t, x_t)$ to the analysis posterior $p_a(\xi_t|y_t, x_{1:t-1})$ at each model time step where retrievals were assimilated. We then calculated a series of other divergences (like in equation (14)) that tested individual EnKF assumptions. We tested four different semiempirical likelihood functions by replacing the data-derived likelihood function in equation (12) with the following:

1. The EnKF observation operator, $\mathscr{N}(y_t|Z_t, R)$, so that the resulting divergence (equation (14)) tested the total information loss due to approximations in the EnKF observation operator.
2. A Gaussian distribution with sample mean and variance: $\mathscr{N}\left(y_t|\widehat{\mu_y}, \widehat{\sigma_y}\right)$, so that the resulting divergence measured information loss due to assuming only that retrieval error has zero higher-order moments.
3. A Gaussian distribution with sample mean and a prescribed variance: $\mathscr{N}\left(y_t|\widehat{\mu_y}, R\right)$, so that the resulting divergence measured information loss due to the prescribed retrieval error covariance.
4. A Gaussian distribution with an identity mean and a sample variance: $\mathscr{N}\left(y_t|Z_t, \widehat{\sigma_y}\right)$, so that the resulting divergence measured information loss due to the linearity assumption in the retrieval operator.

Parameters $\widehat{\mu_y}$ and $\widehat{\sigma_y}$ denote the sample mean and variance of the conditional distribution $p(y_t|Z_t, x_t)$, and R is the (scaled) LPRM retrieval error covariance that was used for DA. Each of these four different likelihood functions resulted in a different posterior distribution, similar to equations (13) and (15.1). Each of the four different posterior distributions that result from using each of the four likelihood functions listed above represents a data-derived (empirical) conditional distribution function that includes some portion of the parametric assumptions used in the likelihood function of the EnKF. By measuring the divergence (as in equation (14)) from the data-derived conditional distribution over Z to these partially parametric conditional distributions over Z allows us to measure the information loss due to (1) assuming that the retrieval is Gaussian distributed with mean given by an identity relationship with the model-simulated soil moisture and variance given by the EnKF ensemble, and that the retrieval is Gaussian distributed with (2) data-derived mean and variance, (3) data-derived mean and prescribed variance, and (4) data-derived variance and prescribed mean.

We tested the same assumptions in the EnKF posterior. The EnKF uses as its prior a Gaussian distribution with mean and variance estimated from the model ensemble at a given time step. We evaluated the effects of using this distribution relative to the data-derived prior using four semiempirical substitutions:

1. The EnKF prior, $\mathscr{N}\left(Z_t|\widehat{X_t}, \widehat{Q_t}\right)$, so that the resulting divergence tested the total information loss due to approximation in the EnKF prior.
2. A Gaussian distribution with data-derived mean and variance: $\mathscr{N}(Z_t|\widehat{\mu_z}, \widehat{\sigma_z})$, so that the resulting divergence measured information loss due to assuming that retrieval error has zero higher-order moments.
3. A Gaussian distribution with sample mean and a prescribed variance: $\mathscr{N}\left(Z_t|\widehat{\mu_z}, \overline{Q_t}\right)$, so that the resulting divergence measured information loss due to assessing model error from the prescribed ensemble characteristics (prescribed forcing errors and state transition errors).
4. A Gaussian distribution with an identity mean and a sample variance: $\mathscr{N}\left(Z_t|\overline{X_t}, \widehat{\sigma_z}\right)$, so that the resulting divergence measured information loss due to the linearity assumption in the retrieval operator.

In the case of these decomposition of the prior, $\widehat{\mu_z}$ and $\widehat{\sigma_z}$ are the sample mean and variance of the conditional distribution $p(Z|X)$. Unlike the efficiency analyses reported in section 3.2, the sample means and variances in this decomposition analysis were derived from the background distributions rather than from the open-loop model simulations, so that we were working with the exact distributions that the EnKF used. The background

Table 4
Divergence Decomposition of the EnKF With LPRM Retrievals According to Equation (14)

Interpretation	Prior	Likelihood	Information loss
Total EnKF divergence	$\mathcal{N}\left[\overline{X}_t, \overline{Q}_t\right]$	$\mathcal{N}[X_t, R]$	4.08
Total effects of model prior	$\mathcal{N}\left[\overline{X}_t, \overline{Q}_t\right]$	$p(y_t \mid x_t, Z_t)$	2.55
Effects of Gaussianity in the prior	$\mathcal{N}\left[\widehat{\mu}_t, \widehat{\sigma}_t\right]$	$p(y_t \mid x_t, Z_t)$	0.06
Effects of ensemble variance	$\mathcal{N}\left[\widehat{\mu}_t, \overline{Q}_t\right]$	$p(y_t \mid x_t, Z_t)$	1.62
Effects of linearity (identity) mean	$\mathcal{N}\left[\overline{X}_t, \widehat{\sigma}_t\right]$	$p(y_t \mid x_t, Z_t)$	0.14
Total effects of retrieval operator[a]	$p(Z_t \mid x_t)$	$\mathcal{N}[Z_t, R]$	1.17
Effects of Gaussianity in retrieval operator	$p(Z_t \mid x_t)$	$\mathcal{N}\left[\widehat{\mu}_t, \widehat{\sigma}_t\right]$	0.08
Effects or prescribed retrieval variance	$p(Z_t \mid x_t)$	$\mathcal{N}\left[\widehat{\mu}_t, R\right]$	1.91
Effects of linearity (identity) mean	$p(Z_t \mid x_t)$	$\mathcal{N}\left[Z_t, \widehat{\sigma}_t\right]$	0.06

Note. These divergences are from the empirical conditional in equation (12) to the specified hybrid conditionals, as described by equation (13). All divergence metrics are reported as a fraction of the entropy of evaluation data, $H(\mathbf{Z})$.
[a]The retrieval operator is often called an *observation operator* in data assimilation literature.

distribution is the prior, $p(X_t \mid x_{1:t-1})$, in equation (10); this differs from the open-loop in that the background distribution at time t includes updates from previous observations.

Divergence results are presented in Table 4. In this case (LPRM assimilation into the Noah-MP model evaluated against the SCAN network), the total EnKF divergence was close to three times the entropy of the in situ observations, and this was split approximately evenly between the likelihood and prior according to equation (16). Within the likelihood itself, almost no divergence was due to the assumption of Gaussian retrieval error, or due to the assumption that the retrieval was unbiased relative to the in situ data. Remember that this lack of retrieval bias relative to the in situ measurements is not site-specific, just that the retrieval in general over the whole run domain is unbiased. Essentially, all of the likelihood-related divergence was due to the retrieval error covariance.

Related to the prior, the divergence was again due predominantly to the model error variance. Again, we used forcing and state perturbations that are essentially standard in soil moisture EnKF applications (e.g., Kumar et al., 2014). Relatively little distributional information about the in situ data was lost due to the Gaussianity assumption, and relatively little was lost due to the assumption the that ensemble mean was unbiased relative to the in situ data. Our model is not biased, we just have not estimated the error variance correctly using what are essentially standard forcing and state transition error terms.

4. Summary and Discussion

The most obvious takeaway from this empirical analysis was that the LPRM space-based soil moisture retrievals that we looked at here contained relatively little information about the point-scale in situ measurements in the SCAN network. This holds regardless of the DA method used to assimilate those observations, and is in general agreement with previous studies that have reported linear-Gaussian metrics with at least comparable values. Even given this low initial information content (related to point-based in situ data), our applications of the EnKF used—at best—only a tiny fraction of the information content of the (imperfect) remote sensing retrievals. It is important to understand that the latter result cannot be extrapolated to other applications of the EnKF to soil moisture DA, since our results here (Table 4) clearly indicate that different choices of observation and state covariance parameters can be expected to improve DA results with even a linear-Gaussian filter like the EnKF.

More generally, we outlined a theory for measuring the ability of DA filters to use the information content of assimilated observations. The primary purpose of this type of procedure is to quantify whether improved DA estimates might come from improved sensor technology versus from improved DA procedures and algorithms. This procedure is generally applicable, and our soil moisture example was just that—an example. It is important to understand that this theory accounts directly for all uncertainties in a typical geophysical data assimilation problem—including model error, retrieval error, and representativeness errors. The latter because information should be preserved through DA even if the model, in situ data, and retrievals are all related to fundamentally different physical quantities.

References

Ball, J. T., Woodrow, I. E., & Berry, J. A. (1987). A model predicting stomatal conductance and its contribution to the control of photosynthesis under different environmental conditions. *Progress in Photosynthesis Research*, 221–224.

Chen, F., Janjić, Z., & Mitchell, K. (1997). Impact of atmospheric surface-layer parameterizations in the new land-surface scheme of the NCEP mesoscale eta model. *Boundary-Layer Meteorology*, *85*(3), 391–421. https://doi.org/10.1023/A:1000531001463

Chen, F., Mitchell, K., Schaake, J., Xue, Y. K., Pan, H. L., Koren, V., et al. (1996). Modeling of land surface evaporation by four schemes and comparison with FIFE observations. *Journal of Geophysical Research*, *101*(D3), 7251–7268. https://doi.org/10.1029/95JD02165

Cover, T. M., & Thomas, J. A. (1991). *Elements of Information Theory*. New York, NY: Wiley-Interscience. https://doi.org/10.1002/0471200611

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

Entekhabi, D., Njoku, E. G., O'Neill, P. E., Kellogg, K. H., Crow, W. T., Edelstein, W. N., et al. (2010). The soil moisture active passive (SMAP) mission. *Proceedings of the IEEE*, *98*(5), 704–716. https://doi.org/10.1109/JPROC.2010.2043918

Evensen, G. (2003). The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, *53*(4), 343–367. https://doi.org/10.1007/s10236-003-0036-9

Jordan, R. (1991). A one-dimensional temperature model for a snow cover: Technical documentation for SNTHERM. 89: DTIC Document.

Kinney, J. B., & Atwal, G. S. (2014). Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(9), 3354–3359.

Knyazikhin, Y., Glassy, J., Privette, J. L., Tian, Y., Lotsch, A., Zhang, Y., et al. (1999). MODIS leaf area index (LAI) and fraction of photosynthetically active radiation absorbed by vegetation (FPAR) product (MOD15). Algorithm theoretical basis document, Version 4 (126 pp.).

Koren, V., Schaake, J., Mitchell, K., Duan, Q.-Y., Chen, F., & Baker, J. M. (1999). A parameterization of snowpack and frozen ground intended for NCEP weather and climate models. *Journal of Geophysical Research*, *104*(D16), 19,569–19,585. https://doi.org/10.1029/1999JD900232

Kumar, S. V., Peters-Lidard, C. D., Mocko, D., Reichle, R., Liu, Y., Arsenault, K. R., et al. (2014). Assimilation of remotely sensed soil moisture and snow depth retrievals for drought estimation. *Journal of Hydrometeorology*, *15*(6), 2446–2469. https://doi.org/10.1175/JHM-D-13-0132.1

Kumar, S. V., Reichle, R. H., Harrison, K. W., Peters-Lidard, C. D., Yatheendradas, S., & Santanello, J. A. (2012). A comparison of methods for a priori bias correction in soil moisture data assimilation. *Water Resources Research*, *48*, W03515. https://doi.org/10.1029/2010WR010261

Kumar, S. V., Reichle, R. H., Peters-Lidard, C. D., Koster, R. D., Zhan, X., Crow, W. T., et al. (2008). A land surface data assimilation framework using the land information system: Description and applications. *Advances in Water Resources*, *31*(11), 1419–1432. https://doi.org/10.1016/j.advwatres.2008.01.013

Liu, Q., Reichle, R. H., Bindlish, R., Cosh, M. H., Crow, W. T., de Jeu, R., et al. (2011). The contributions of precipitation and soil moisture observations to the skill of soil moisture estimates in a land data assimilation system. *Journal of Hydrometeorology*, *12*(5), 750–765. https://doi.org/10.1175/JHM-D-10-05000.1

Maggioni, V., & Houser, P. R. (2017). Soil moisture data assimilation. In S. K. Park & L. Xu (Eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications* (Vol. 3, pp. 195–217). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-43415-5_9

Maggioni, V., Reichle, R. H., & Anagnostou, E. N. (2011). The effect of satellite rainfall error modeling on soil moisture prediction uncertainty. *Journal of Hydrometeorology*, *12*(3), 413–428. https://doi.org/10.1175/2011JHM1355.1

Maggioni, V., Reichle, R. H., & Anagnostou, E. N. (2013). The efficiency of assimilating satellite soil moisture retrievals in a land data assimilation system using different rainfall error models. *Journal of Hydrometeorology*, *14*(1), 368–374. https://doi.org/10.1175/JHM-D-12-0105.1

McNally, A., Arsenault, K., Kumar, S., Shukla, S., Peterson, P., Wang, S., et al. (2017). A land data assimilation system for sub-Saharan Africa food and water security applications. *Scientific Data*, *4*, 170012. https://doi.org/10.1038/sdata.2017.12

Mladenova, I. E., Nearing, G. S., Bolten, J. D., & Lakshmi, V. (2017). Remote sensing techniques and data assimilation for hydrologic modeling. In V. Singh (Ed.), *Handbook of Applied Hydrology* (2nd ed., Chapter 8). McGraw-Hill.

Moran, M. S., Doorn, B., Escobar, V., & Brown, M. E. (2015). Connecting NASA science and engineering with earth science applications. *Journal of Hydrometeorology*, *16*(1), 473–483. https://doi.org/10.1175/JHM-D-14-0093.1

Nearing, G. S., Gupta, H. V., Crow, W. T., & Gong, W. (2013). An approach to quantifying the efficiency of a Bayesian filter. *Water Resources Research*, *49*, 2164–2173. https://doi.org/10.1002/wrcr.20177

Niu, G. Y., & Yang, Z. L. (2004). Effects of vegetation canopy processes on snow surface energy and mass balances. *Journal of Geophysical Research*, *109*, D23111. https://doi.org/10.1029/2004JD004884

Niu, G.-Y., & Yang, Z.-L. (2006). Effects of frozen soil on snowmelt runoff and soil water storage at a continental scale. *Journal of Hydrometeorology*, *7*(5), 937–952. https://doi.org/10.1175/JHM538.1

Niu, G. Y., Yang, Z. L., Dickinson, R. E., & Gulden, L. E. (2005). A simple TOPMODEL-based runoff parameterization (SIMTOP) for use in global climate models. *Journal of Geophysical Research*, *110*, D21106. https://doi.org/10.1029/2005JD006111

Niu, G. Y., Yang, Z. L., Dickinson, R. E., Gulden, L. E., & Su, H. (2007). Development of a simple groundwater model for use in climate models and evaluation with gravity recovery and climate experiment data. *Journal of Geophysical Research*, *112*, D07103. https://doi.org/10.1029/2006JD007522

Niu, G. Y., Yang, Z. L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., et al. (2011). The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *Journal of Geophysical Research*, *116*, D12109. https://doi.org/10.1029/2010JD015139

Oleson, K. W., Lawrence, D. M., Gordon, B., Flanner, M. G., Kluzek, E., Peter, J., et al. (2010) Technical description of version 4.0 of the Community Land Model (CLM).

Owe, M., de Jeu, R., & Holmes, T. (2008). Multisensor historical climatology of satellite-derived global land surface moisture. *Journal of Geophysical Research*, *113*, F01002. https://doi.org/10.1029/2007JF000769

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Computation*, *15*(6), 1191–1253. https://doi.org/10.1162/089976603321780272

Reichle, R., Crow, W., Koster, R., Kimball, J., & De Lannoy, G. (2016). SMAP Level 4 Surface and Root Zone Soil Moisture (L4_SM) Data Product.

Rodell, M., Houser, P., Jambor, U. E. A., Gottschalck, J., Mitchell, K., Meng, C., et al. (2004). The global land data assimilation system. *Bulletin of the American Meteorological Society*, *85*(3), 381–394. https://doi.org/10.1175/BAMS-85-3-381

Schaake, J. C., Koren, V. I., Duan, Q.-Y., Mitchell, K., & Chen, F. (1996). Simple water balance model for estimating runoff at different spatial and temporal scales. *Journal of Geophysical Research*, *101*(D3), 7461–7475. https://doi.org/10.1029/95JD02892

Schaefer, G. L., Cosh, M. H., & Jackson, T. J. (2007). The USDA natural resources conservation service soil climate analysis network (SCAN). *Journal of Atmospheric and Oceanic Technology*, *24*(12), 2073–2077. https://doi.org/10.1175/2007JTECHA930.1

Schneidman, E., Bialek, W., & Berry, M. J. (2003). Synergy, redundancy, and independence in population codes. *The Journal of Neuroscience*, *23*(37), 11,539–11,553.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

van Leeuwen, P. J. (2010). Nonlinear data assimilation in geosciences: An extremely efficient particle filter. *Quarterly Journal of the Royal Meteorological Society*, *136*(653), 1991–1999. https://doi.org/10.1002/qj.699

Verseghy, D. L. (1991). CLASS—A Canadian land surface scheme for GCMs. I. Soil model. *International Journal of Climatology*, *11*(2), 111–133.

Wikle, C. K., & Berliner, L. M. (2007). A Bayesian tutorial for data assimilation. *Physica D: Nonlinear Phenomena*, *230*(1–2), 1–16. https://doi.org/10.1016/j.physd.2006.09.017

Xia, Y., Cosgrove, B. A., Ek, M. B., Sheffield, J., Luo, L., Wood, E. F., et al. (2013). Overview of the North American Land Data Assimilation System (NLDAS), Land Surface Observation, Modeling and Data Assimilation: World Scientific. (pp. 337–377).

Xia, Y., Ek, M., Wood, E., Sheffield, J., Luo, L., & Mo, K. (2011). North American land data assimilation system (NLDAS) in support of the U.S. drought and flood analysis, monitoring, and prediction, including National Integrated Drought Information System (NIDIS). International Union of Geodesy and Geophysics, Melbourne, Australia.

Xue, Y., Sellers, P. J., Kinter, J. L., & Shukla, J. (1991). A simplified biosphere model for global climate studies. *Journal of Climate*, *4*(3), 345–364. https://doi.org/10.1175/1520-0442(1991)004<0345:ASBMFG>2.0.CO;2

Yang, Z.-L., & Dickinson, R. E. (1996). Description of the biosphere-atmosphere transfer scheme (BATS) for the soil moisture workshop and evaluation of its performance. *Global and Planetary Change*, *13*(1-4), 117–134. https://doi.org/10.1016/0921-8181(95)00041-0

Yang, Z.-L., Dickinson, R. E., Robock, A., & Vinnikov, K. Y. (1997). Validation of the snow submodel of the biosphere-atmosphere transfer scheme with Russian snow cover and meteorological observational data. *Journal of Climate*, *10*(2), 353–373. https://doi.org/10.1175/1520-0442(1997)010<0353:VOTSSO>2.0.CO;2