Nonparametric Triple Collocation

Grey Nearing – University of Alabama
et al.

Citation of published version:

Nearing, G., et al. (2017): Nonparametric Triple Collocation. *Water Resources Research*, 53. DOI: 10.1002/2017WR020359

AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2017WR020359

Nonparametric triple collocation

Grey S. Nearing[1,2,3], Soni Yatheendradas[1,4], Wade T. Crow[5], David D. Bosch[6], Michael H. Cosh[5], David C. Goodrich[7], Mark S. Seyfried[8], and Patrick J. Starks[9]

[1]NASA-GSFC, Hydrologic Sciences Laboratory, Greenbelt, Maryland, USA, [2]Science Systems International Corporation, Greenbelt, Maryland, USA, [3]NCAR, Research Applications Laboratory, Boulder, Colorado, USA, [4]Earth Science Systems Interdisciplinary Center, University of Maryland, College Park, Maryland, USA, [5]USDA-ARS, Hydrology and Remote Sensing Laboratory, Beltsville, Maryland, USA, [6]USDA-ARS, Southeast Watershed Research Laboratory, Tifton, Georgia, USA, [7]USDA-ARS, Southwest Watershed Research Center, Tucson, Arizona, USA, [8]USDA-ARS, Grazinglands Research Laboratory, El Reno, Oklahoma, USA, [9]USDA-ARS, El Reno, Oklahoma, USA

Key Points:
- We present a nonlinear version of triple collocation
- Linear assumptions in standard triple collocation result in information loss and corruption
- Nonlinear triple collocation appears to perform better than linear triple collocation in our soil moisture example

Correspondence to:
G. S. Nearing,
grey.s.nearing@nasa.gov

Citation:
Nearing, G. S., S. Yatheendradas, W. T. Crow, D. D. Bosch, M. H. Cosh, D. C. Goodrich, M. S. Seyfried, and P. J. Starks (2017), Nonparametric triple collocation, *Water Resour. Res.*, *53*, 5516–5530, doi:10.1002/2017WR020359.

Received 1 JAN 2017
Accepted 11 MAY 2017
Accepted article online 17 MAY 2017
Published online 7 JUL 2017

Abstract Triple collocation has found widespread application in the hydrological sciences because it provides information about the errors in our measurements without requiring that we have any direct access to the true value of the variable being measured. Triple collocation derives variance-covariance relationships between three or more independent measurement sources and an indirectly observed truth variable in the case where the measurement operators are additive. We generalize that theory to arbitrary observation operators by deriving nonparametric analogues to the total error and total correlation statistics as integrations of divergences from conditional to marginal probability ratios. The nonparametric solution to the full measurement problem is underdetermined, and we therefore retrieve conservative bounds on the theoretical total nonparametric error and correlation statistics. We examine the application of both linear and nonlinear triple collocation to synthetic examples and to a real-data test case related to evaluating space-borne soil moisture retrievals using sparse monitoring networks and dynamical process models.

1. Introduction

1.1. Motivation

Triple collocation (TC) is a technique used to estimate the second-order statistics of independent additive errors on three [*Stoffelen*, 1998] or more [*Scipal et al.*, 2010] measurement sources without any explicit requirement to know the true values of the measured quantities. This is quite powerful in situations where we want to understand data error but do not have any data source that we consider accurate enough to represent reference values for the quantity of interest. An accessible overview of TC is given by *Vogelzang and Stoffelen* [2012], and reviews of recent applications across the geoscience are given by *McColl et al.* [2014] and *Gruber et al.* [2016].

One of the most common applications of TC is to estimate the error structures of remote sensing data products when it is not feasible to collect in situ data at the scale of a satellite pixel [e.g., *Scipal et al.*, 2010; *Chen et al.*, 2017; *Yilmaz and Crow*, 2014; *Miyaoka et al.*, 2017; *Roebeling et al.*, 2012]. TC requires that all measurement sources are related to the "true" signal up to a scaling constant plus some additive error that is independent of truth. In many applications, this assumption is violated a priori simply by the fact that scale mismatches mean that remote sensing retrievals and in situ instruments means do not actually observe the same physical quantity [*Gruber et al.*, 2016].

More generally, *Gruber et al.* [2016] identified five necessary assumptions about the structure of the measurement sources required by TC:

1. **Linearity**: that all measurement sources are a combination of the scaled value of the true variable of interest plus some additive random noise.
2. **Independence of signal and noise**: that the error in each measurement source is independent of the true quantity of interest.
3. **Independence across measurements**: that the noise distributions are independent across the measurement sources.

4. **Stationarity**: that the signal and error properties of the available measurement samples are representative, and therefore can be extrapolated.
5. **Representativeness**: that the three (or more) measurement sources all measure essentially the same system property.

Those same authors [*Gruber et al*., 2016] point out that these assumptions are difficult to justify or asses in practice, and that "[w]hile many studies have investigated issues related to possible violations of the underlying assumptions, only few TC modifications have been proposed to mitigate the impact of these violations."

We develop here a nonparametric interpretation of the three-measurement problem that alleviates the need for three of these five assumptions (items 1, 2, and 5). In particular, we propose a fully nonparametric form of TC that treats the full joint probability distribution between all measurements and the (unobserved) truth. The result is that it is not necessary to assume any particular form of the measurement operator—in particular, it is not necessary to assume that the measurement noise is additive or independent of the true value of the observed variable. We require only that there be a joint distribution between some true quantity and all three measurements—not even that each of the three measurements "observe" the same quantity directly. We do, however, see no way of alleviating the assumption that the individual measurement sources have independent errors, and, of course, the stationarity assumption is inherent in any application of either probability theory or statistics.

Given that most real-world dynamical systems contain nonlinearities like missed events, scale-dependent noise, power law effects, thresholds, etc., it seems reasonable to assume that measurements of these systems will often contain strong nonlinearities. As such, we expect that it would be beneficial to understand how to treat nonlinear measurement errors without imposing any a priori assumptions about the parametric form of the underlying probability measurement distributions.

The arbitrary situation is underdetermined. What we do in this essay is highlight: (i) what *can* and *cannot* be known about the errors of three coincident independent measurement sources absent strong linearity assumptions, and (ii) the sense in which strong linearity assumptions force the values of the underdetermined aspects of the problem. The latter is demonstrated both theoretically and empirically.

1.2. Background

Triple collocation works as follows. Suppose a true state $t=\{t_n\}_{n=1,\ldots,N}$ is observed indirectly by three or more independent measurement sources (denoted by subscripts i, j, k) that each take the form:

$$x_{i,n} \sim f_i\left(X_{i,n}|T_n=t_n\right). \tag{1}$$

Our notation is such that capital letters represent random variables, lower-case letters represent realizations of random variables, the first index on measurements represents the measurement source, and the second (or only) index represents a sample from that source. In the case where the measurements from each source are stationary and iid, and the source operators are additive and linear in the mean we have:

$$x_{i,n}=\beta_i t_n+\varepsilon_{i,n}. \tag{2}$$

To a second-order approximation, this is:

$$x_{i,n} \sim \mathcal{N}\left(\beta_i t_n, \sigma_i^2\right). \tag{3}$$

The actual noise distributions in the measurement sources may be arbitrary and must obey only weak constraints like homoscedasticity and finite moments, however the standard methods of TC consider only the second moments of these distributions.

The second-order quantities that are estimable directly from sample measurements are $c_{ij}=cov\left(X_i, X_j\right)$ and $c_{ii}=var\left(X_i\right)$. We may use independence between the measurement sources to derive the error variances of each source from these as [*Caires and Sterl*, 2003]:

$$\sigma_i^2=c_{ii}-\frac{c_{ij}c_{ik}}{c_{jk}}\ . \tag{4}$$

We also may derive the correlations between the measurement sources and truth as [*McColl et al*., 2014]:

$$\rho_{i,t}^2 = \frac{c_{ij}c_{ik}}{c_{ii}c_{jk}}, \tag{5}$$

and also a scaled estimate of the variance of the truth:

$$\sigma_t^2 = \frac{c_{ij}c_{ik}}{\beta_i^2 c_{jk}}. \tag{6}$$

Equation (6) is symmetric in the measurement sources.

Equations (4)–(6) are powerful and have found widespread application in the hydrological sciences precisely because they tell us something about the errors in our measurements without requiring that we have any direct access to the true value of the thing being measured. In particular, these equations return second-order properties of the noise (equation (4)) and signal (equation (5)) in coincident measurement sources.

2. Generalization

2.1. Measuring Information

The problem is to choose a set of metrics that represents meaningful integrations of the various joint and conditional probability distributions among $\{X_i, X_j, X_k, T\}$. That is, our desideratum is a set of statistics over ratios of arbitrary joint and conditional probability distributions that fully measure the dependence and independence of the various measurement sources and the truth.

To motivate our choice of metrics, consider how a rational knowledge state evolves against evidence. We work in the context of *Cox*' [1946] theorem, which shows that probability theory is a logical calculus. Before collecting a set of measurements, any distributive measure over our proposition set scales multiplicatively across the number of data according to a product rule. For example, the probability we place on any specific outcome of two independent measurements is the product of the marginal probabilities of those outcomes. On the other hand, after collecting measurements, our proposition set scales additively according to a sum rule: to summarize everything that we now know about the system we need an amount of information that is proportional to the number of measurement data, less any effects of nonindependence. This means that the appropriate model of the epistemic consequences of collecting a set of measurements is described fully by the axioms of *Shannon*'s [1948] second theorem. *Knuth* [2005] gives a similar uniqueness theorem in the context of logic of measurement.

The implication is that in context of probability theory we have essentially only one choice of metric that may be used to quantify the expected information content of measurements. In particular, we quantify ignorance as the expected change in our knowledge state that would occur if we were able to measure truth directly:

$$H(T) = E\left[\ln\left(p(T)^{-1}\right)\right]. \tag{7}$$

H is sometimes called *entropy*. Similarly, the information provided by any indirect measurement source X_i is the expected reduction in ignorance due to conditioning on that source:

$$I(T;X_i) = E\left[\ln\left(\frac{p(T|X_i)}{p(T)}\right)\right]. \tag{8}$$

$$I(T;X_i) = H(T) - H(T|X_i) = H(X_i) - H(X_i|T) \tag{9}$$

$I(T;X)$ is referred to as the *mutual information* between random variables T and X. The quantity $H(T) - H(T|X_i)$ is interpreted as the amount of variability in T that can be accounted for by a single measurement source X_i, while the quantity $H(X_i) - H(X_i|T)$ is interpreted as the variability in X_i that is signal rather than noise. These two quantities are always equal because the expected value in equation (8) integrates over the joint distribution, $p(T, X_i)$, and because the probability ratio in log term may be replaced by: $\frac{p(T|X_i)}{p(T)} = \frac{p(T,X_i)}{p(T)p(X_i)} = \frac{p(X_i|T)}{p(X_i)}$.

The significance of these metrics with respect to the problem of understanding error characteristics of measurement sources is as follows:

1. The ratio $H(X_i|T)/H(X_i)$ is the fraction of variability in the measurement source that is due to error. This quantifies bad information (error) in the measurement and is analogous (up to scaling) with equation (4). In actuality, $H(X_i|T)$ is directly analogous to σ_i^2, but we prefer to use scaled metrics.
2. The ratio $I(T;X_i)/H(X_i)$ is the normalized nonparametric correlation between the measurement source and the truth. This quantifies good information in the measurement and is analogous (up to scaling) with equation (5).
3. The ratio $H(T|X_i)/H(T)$ is the fraction of information about the truth that is not contained in the measurement source X_i. This quantifies information missing from the measurement and has no analogue in equations (4)–(6).

These ratios are the statistics that we would like to estimate, and if our variables are discrete then each ratio is bounded between 0 and 1 with the usual implications of independence or identity at the extremes [*Paninski*, 2003].

These basic information concepts may be extended to any number of variables. For example, $I(T;X_i,X_j)$ is the amount of information about the truth that is contained in two measurement sources, and $H(T|X_i,X_j)/H(T)$ is the fraction of information about the truth that is not captured by either measurement source. Interestingly, multivariate information statistics, e.g., $I(T;X_i;X_j)$, may take on either a negative or positive value, with a positive value indicating that there is some redundancy in information provided by the two measurement sources while a negative value indicates that there is synergy between the two sources [*Schneidman et al*., 2003] (Synergy explained: In any system with one random variable, we have entropy according to equation (7). In any system with two or more random variables, one variable may contain information about the other according to equation (9). In a system with three or more random variables, any pair of variables *may* carry a total amount of information about the third variable that is greater than the sum of the information contents of the two predictors separately—this is synergy. A simple example is the summation of two fair dice, which has entropy H = 2.27 nats. Since both dice together completely determine the sum we know that the information carried by the two dice about the sum is equal to the entropy of the sum, however the mutual information between any one die and the sum is only 0.48 nats, which is less than half of the total information.)

2.2. Nonparametric Triple Collocation

We cannot directly estimate any of the three desired information ratios outlined in section 2.1 because we do not have access to the true states of the system and we therefore cannot derive empirical joint distributions like $p(T,X_i)$. Instead, we may estimate only the shared information between the three sets of measurements—both pairwise and in the triplet.

Figure 1 illustrates the most general expression of the TC problem that recognizes epistemic uncertainty (i.e., uncertainty about the form of the measurement operators); the four primary objects in this Venn diagram represent the entropy of the three measurement sources and of the truth variable according to equation (7), and the overlapping sections represent the information that the various variables share according to equations (8) and (9). Since any three-way or four-way shared information—e.g., $I(X_i;X_j;X_k)$—may be negative due to synergy, this Venn diagram is over signed, additive measures. The quantities that are necessary to calculate our desired statistics are shaded: we have direct access to estimators of the information spaces shaded purple, we do not have access to the information spaces shaded pink, and the areas shaded gray are zero under an assumption that all measurements are independent conditional on the truth. In addition to the quantities shaded purple in Figure 1, we may also directly estimate entropies of all three measurement sources; e.g., $H(X_i)$.

Assuming independence in the sources conditional on the truth, then the conditional mutual information metrics completely measure *everything that we can know* (in the expectation) about missing information and bad information or disinformation in each measurement source:

1. $H(X_i|X_j,X_k)$ measures everything that we can know about the total error in source X_i.
2. 2. $I(X_i;X_j)+I(X_i;X_k|X_j)$ measures everything that we can know about the total (good) information in source X_i.
3. 3. $I(X_k;X_j|X_i)$ measures everything that we can know about the information missing from source X_i.

These quantities are illustrated in Figure 2.

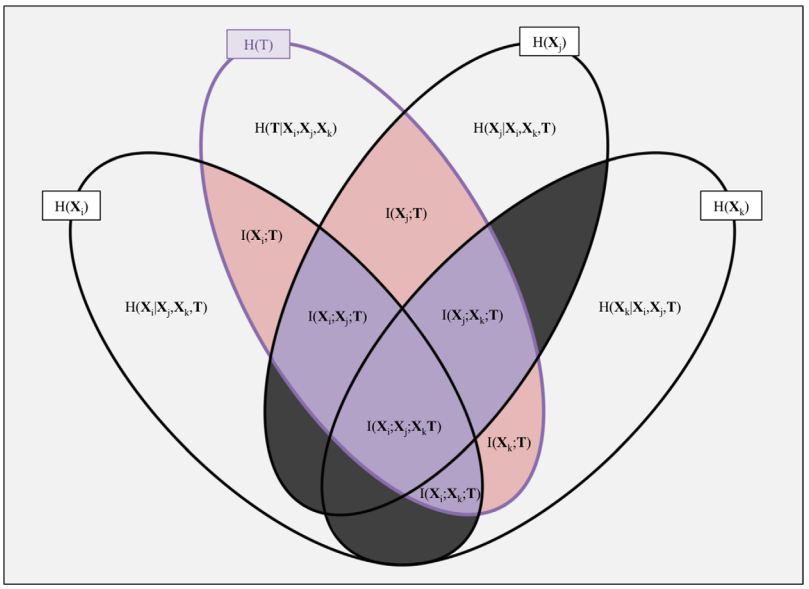

Figure 1. This information diagram illustrates a general statement of the TC problem. The four primary components are the entropies of the three measurement sources and the truth according to equation (4), and the overlapping areas represent information that is shared between variables according to equations (8) and (9). The gray-shaded regions are zero under the assumption that the measurements are independent conditional on the truth. The purple regions represent portions of the problem that can be estimated directly from available measurement samples, and the pink-shaded regions represent the portions of the problem that we would need to know to directly estimate total error and total information (either linear or nonlinear). The pink-shaded regions are the portions of the information space that we cannot measure, and these account for the inequalities in equations (13) and (14).

Again, we will normalize the total error and total information by $H(X_i)$ so that that in the discrete case these quantities range between 0 and 1. The missing information quantity can also be normalized by $H(X_i)$ so as to be directly comparable with the other two, however notice that this is not necessarily bounded at unity; it would be bounded above if we could normalize by $H(T)$, but we cannot since we cannot know the entropy of a random variable representing truth.

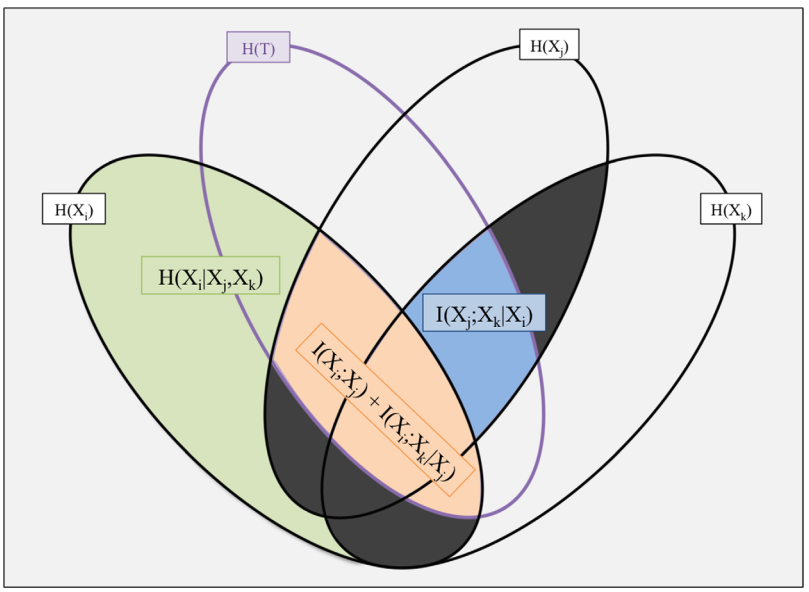

Figure 2. The same Venn diagram as Figure 1, except that here the colored sections represent those measurable quantities outlined in section 2.1. $H(X_i|X_j,X_k)$ measures everything that we can know about the total error in source X_i. $I(X_i;X_j)+I(X_i;X_k|X_j)$ measures everything that we can know about the total (good) information in source X_i. $I(X_k;X_j|X_i)$ measures everything that we can know about the information missing from source X_i.

We can see from Figure 1 that the desired statistics outlined in section 2.1 are:

$$H(X_i|T)=H(X_i|X_j,X_k)-I(T;X_i|X_j,X_k), \tag{10}$$

$$I(T;X_i)= I(X_i;X_j)+I(X_i;X_k|X_j)+I(T;X_i|X_j,X_k), \tag{11}$$

$$H(T|X_i)= I(X_k;X_j|X_i)+I(T;X_j|X_i,X_k)+I(T;X_k|X_i,X_j)+H(T|X_i,X_j,X_k). \tag{12}$$

These are bounded, by removal of certain positive terms, by the measurable portions of the problem as:

$$H(X_i|T) \leq H(X_i|X_j,X_k), \tag{13}$$

$$I(T;X_i) \geq I(X_i;X_j)+I(X_i;X_k|X_j), \tag{14}$$

$$H(T|X_i) \geq I(X_k;X_j|X_i). \tag{15}$$

Notice that the first two bounds—on total error and total information—are conservative. The nonequality in the total error and total correlation bounds (equations (13) and (14)) is due to the underdetermined correlations: $I(T;X_i|X_j,X_k)$, while the nonequality in the missing information statistic (equation (15)) is due to underdetermined correlation plus fundamentally unobserved variability in the truth. Our objective in this paper is to understand how the consequences of bounding rather than estimating nonparametric total error statistics compare with the consequences of imposing strong linearity assumptions and ignoring higher-order moments.

These concepts can, at least in principle, be extended to any number of sources: i.e., via $I(X_{\sim i}|X_i)$ and $H(X_i|X_{\sim i})$. However, in practice such sample statistics become increasingly difficult to estimate due to the problem of dimensionality when estimating empirical joint distributions.

2.3. Linear Measurements as a Special Case

Unlike the variance-covariance metrics in equation (4), the information-theoretic metrics are not absolute—they represent everything that we can learn about the error properties of three or more stationary measurement sources from a set of measurement samples. The linearity assumptions in equation (2) coupled with the implicit treatment of all error sources as Gaussian (standard TC does not require this, but it does ignore all higher-order moments) are equivalent to assuming certain properties about each of the sections of the Venn diagram in Figure 1, including about the fundamentally unmeasurable quantities $I(T;X_i|X_j,X_k)$. Our position is that it is exactly the extent to which we make a priori assumptions about the information characteristics of our data that we make errors in understanding the *actual* information characteristics of our data.

Relationships between nonparametric TC statistics and the measurement distributions can be derived exactly for continuous random variables under the assumption that all higher-order moments are zero:

$$I(X_i;X_j)=-\frac{1}{2}\ln\left(1-\frac{c_{ij}^2}{c_{ii}c_{jj}}\right), \tag{16}$$

$$I(X_i;X_j|X_k)=\frac{1}{2}\ln\left(\frac{\left(c_{ii}-\frac{c_{ik}^2}{c_{kk}}\cdot\right)\left(c_{jj}-\frac{c_{jk}^2}{c_{kk}}\cdot\right)}{-c_{ij}^2+c_{ii}c_{jj}-\frac{c_{ik}^2c_{jj}}{c_{kk}}+\frac{2c_{ij}c_{ik}c_{jk}}{c_{kk}}-\frac{c_{jk}^2c_{ii}}{c_{kk}}}\right) \tag{17}$$

$$H(X_i|T)=\frac{1}{2}\ln\left(2\pi e\sigma_i^2\right)=\frac{1}{2}\ln\left(2\pi e\left(c_{ii}-\frac{c_{ij}c_{ik}}{c_{jk}}\right)\right), \tag{18}$$

$$I(X_i;T)=-\frac{1}{2}\ln\left(1-\rho_{i,t}^2\right)=-\frac{1}{2}\ln\left(1-\frac{c_{ij}c_{ik}}{c_{ii}c_{jk}}\right). \tag{19}$$

Notice that equations (16) and (17) are calculable directly from observables, and we can thus measure whether the linearity assumptions and second-order restrictions cost us anything here during any particular TC application. For example, we can measure $I(X_i;X_j)$ directly and also (independently) calculate the right-hand side of equation (16) directly. Any difference between these two values is due to a combination of the linearity assumption in equation (2) plus the fact that linear TC only treats second moments of the error distributions. We can do the same thing for equation (17), however equations (18) and (19) are different. The

latter two describe the relationship between the parametric and nonparametric total error and total correlation (total information) statistics, which are our ultimate objectives. Here we cannot know the value of the quantity $I(T;X_i|X_j,X_k)$ without knowing the exact form of the measurement operators up to their full probability distributions. These last two relationships are thus purely illustrative of how the information theoretic values are forced (at least up to their maximum-entropy approximations) by the strong assumption in equation (2) and approximation in equation (3).

We may go the other direction as well. Even though we cannot estimate either $I(T;X_i)$ or $H(X_i|T)$ directly in the general case, equations (16)–(19) do allow us to solve for the variance-covariance metrics equations (4)–(6) uniquely under the assumption that equation (3) holds—i.e., that the measurement errors are *actually* Gaussian; these are:

$$\sigma_i^2 = \frac{1}{2e\pi}\left[\left(\frac{\left(4e^2\pi^2 - e^{2H(X_i,X_j)}\right)\left(4e^2\pi^2 - e^{2H(X_i,X_k)}\right)}{\left(4e^2\pi^2 - e^{2H(X_j,X_k)}\right)}\right)^{\frac{1}{2}} + e^{2H(X_i)}\right] \tag{20}$$

$$\rho_{i,t}^2 = e^{-2(X_i)}\left(\frac{\left[4e^2\pi^2 - e^{2H(X_i,X_j)}\right]\left[4e^2\pi^2 - e^{2H(X_i,X_k)}\right]}{\left[4e^2\pi^2 - e^{2H(X_j,X_k)}\right]}\right)^{\frac{1}{2}}. \tag{21}$$

Finally, let us establish an intuition about the basic relationship between total information and total error ratios and linear-Gaussian noise in general. Figure 3 shows how the normalized mutual information $I(X_i;T)/H(X_i)$ shown on the x axis changes as a function of the variance of the (linear) measurement noise σ_i^2 and also with the correlation coefficient, $\rho_{i,t}^2$ assuming that equation (2) holds. This illustrates the basic relationship described by equation (19), but calculated numerically for discrete entropy and information metrics. Figure 3 also shows the same thing for equation (18): how the total error fraction $H(X_i|T)/H(X_i)$ varies with σ_i^2 and $\rho_{i,t}^2$ under Gaussian errors.

3. Examples

In this section we present several examples of deriving both linear and nonlinear TC statistics. Three of these examples are synthetic— one synthetic example is truly linear according to equation (2), one is linear but with non-Gaussian errors, and one is nonlinear. In these synthetic cases the truth is known exactly. The intent of these synthetic examples is to highlight differences and trade-offs between parametric and nonparametric strategies for using collocated measurements. The real-data example is to use several "dense" in situ networks to understand the reliability of using a sparse in situ network for evaluating soil moisture remote sensing products. Soil moisture remote sensing evaluation is one of the primary uses of linear TC.

In each example we measured the "true" values of the four primary quantities of interest: $H(X_i|T)/H(X_i)$, $I(X_i;T)/H(X_i)$, σ_i^2, and $\rho_{i,t}^2$, as well as the estimates of each of these quantities using either linear or nonlinear TC as appropriate. In the real-data examples we took as truth spatial averages of precipitation or soil moisture from dense in situ networks, and then showed how using a sparse validation network affects the reliability of the results.

3.1. Synthetic Examples

3.1.1. Linear Example

Our first synthetic example demonstrates that the nonparametric analysis returns reasonable estimates in the case where the linear measurement operators hold exactly. In this example we set all $\beta_i = \beta_j = \beta_k = 1$ and varied the noise standard deviations between $0 \leq \sigma \leq 0.5$. All measurement sources had the same error variance. The truth was a set of draws from a standard normal distribution: $t_{1:n} \sim \mathcal{N}(0,1)$. All information statistics were calculated using discrete maximum likelihood estimators [*Paninski*, 2003] with bin widths ranging from 2% to 20% of the total range of measurement data. Because this is a nonparametric method that requires sufficient data to estimate empirical joint probability distributions, we used sample size that ranged from $N = 10^2$ to $N = 10^5$.

Figure 4 shows results for a sample size of $N = 10^4$. The nonlinear bounds on total error and total information are very close to the true estimates (the differences are due completely to the inequalities in equations

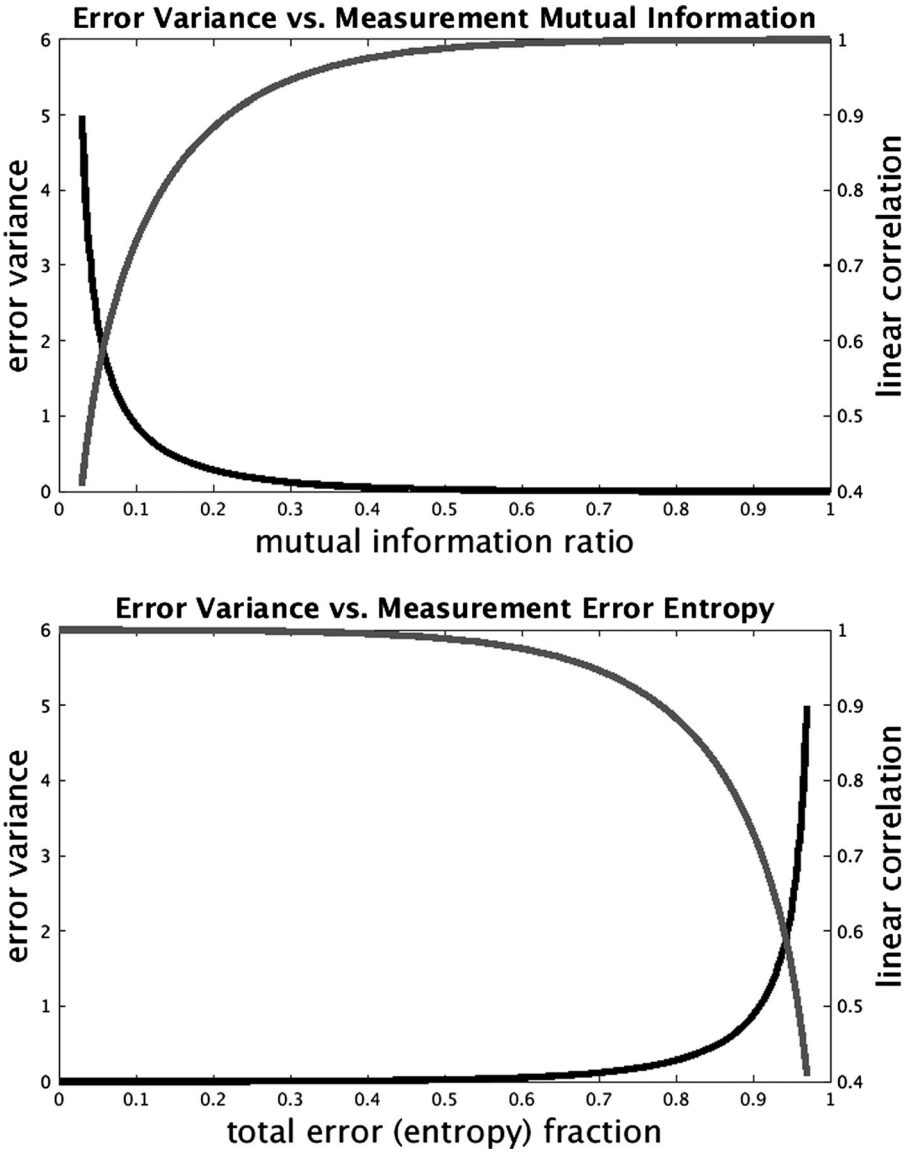

Figure 3. Theoretical relationships between (top) total nonlinear correlation and (bottom) total nonlinear error statistics on the *x* axes versus the linear analogues (on the *y* axes) obtained from traditional TC in the case where the linearity assumptions hold exactly and the information statistics are calculated numerically using discretized random variables.

(13)–(15), however we cannot reliably estimate missing information. Of course, the linear statistics are exactly correct and are therefore not shown. In this particular case, the choice of bin width (different colored lines in the plots) does not have a strong effect on results.

Figure 5 shows convergence of the estimators with increasing sample size. In this ideal case, linear triple colocation requires about $N=10^3$ data points to achieve stability, whereas the nonparametric estimators require an order of magnitude more data to achieve stability at a resolution of about 5% of the measurement range. Convergence rates of the nonparametric estimators depend strongly on the resolution of the empirical joint density functions.

3.1.2. Non-Gaussian Example

Our second synthetic example explores the effect of non-Gaussian error sources on linear and nonparametric TC. We applied both techniques to measurements with additive errors from log-normal and generalized extreme value (GEV) distributions. In the case of GEV errors, we performed several experiments by letting the shape

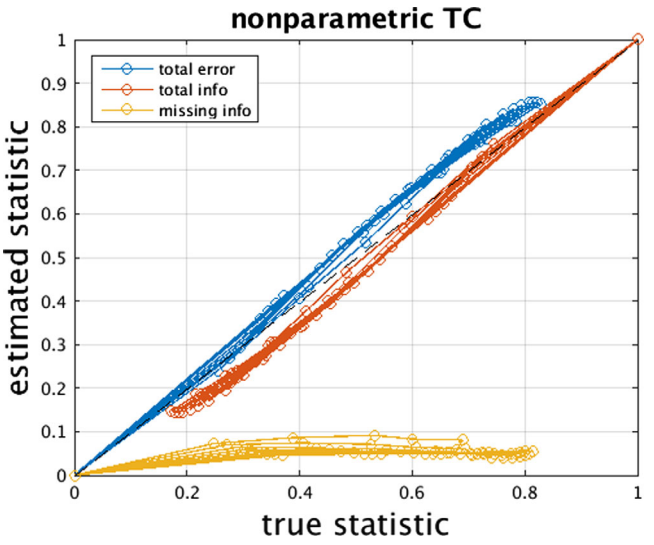

Figure 4. Results from a nonparametric TC analysis of the situation where all measurement sources are subject to additive Gaussian noise. Different lines of each color represent different bin-widths used for discrete entropy and information estimators—these bin widths range from 1% to 20% of the total range of the measurement data. Different points on each line represent different noise standard deviations with $0 \leq \sigma \leq 0.5$. All measurement sources have the same variance in each experiment.

parameter vary from $-2 \leq k \leq 1$, and the scale parameter vary between $0 \leq s < 1$. In the case of the log-normal distribution, we performed several experiments by letting the scale parameter range from $0 \leq s \leq 1$.

In all cases, we also generated Gaussian random errors with the same variances as the log-normal or GEV errors. As expected, the linear TC statistics were identical over Gaussian and non-Gaussian errors with identical variances. Since these are all ideal cases in the sense that the measurements obey $x_{i,n}=t_n+\varepsilon_{i,n}$, linear TC returns essentially the exact variance of the additive errors $\varepsilon_{i,n}$. That being said, linear TC fails to account for any non-Gaussianity in the error structure—the results are the same for Gaussian versus non-Gaussian errors with similar variances.

Nonparametric TC does account for non-Gaussianity in the error structure, but again returns only bounded estimators. This is shown in Figure 6, which compares the absolute difference between the true Gaussian versus non-Gaussian error statistics, $\frac{H(X_i|T)}{H(X_i)}$, with the difference between estimated versus true statistics: i.e., $\frac{H(X_i|X_j,X_k)}{H(X_i)} - \frac{H(X_i|T)}{H(X_i)}$. Figure 6 plots results only for a bin-width of 4% of the total range of measurement data, but results for other bin-widths (ranging from 1% to 20% of the total range of measurement data) were almost identical. What we see in these two cases is that there is a tradeoff between estimator error and error due to only considering the second moment of the error distribution. In most cases that we have found, the latter is a larger effect that the former.

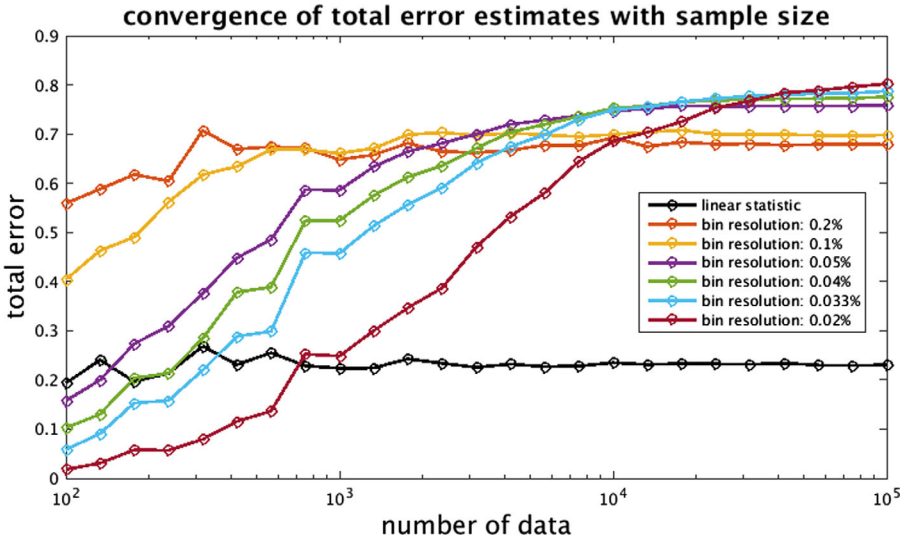

Figure 5. Nonparametric total error statistics require at least an order of magnitude more data to achieve stability than do their linear counterparts.

3.1.3. Nonlinear Example

Our third synthetic example demonstrates both linear and nonlinear TC statistics applied to a nonlinear measurement problem. Here we again took *iid* standard normal samples as the truth and structured the measurement sources as follows:

$$x_{i,n}=t_n+\varepsilon_{i,n}\ ;\ \varepsilon_{i,n}\sim\mathcal{N}(0,\sigma^2)\ \leftrightarrow x_{i,n}\sim\mathcal{N}(t_n,\sigma^2), \tag{22}$$

$$x_{j,n}=\varepsilon_{j,n}t_n\ ;\ \varepsilon_{j,n}\sim\mathcal{N}(0,\sigma^2)\ \leftrightarrow\ x_{j,n}\sim\mathcal{N}(0,t_n^2\sigma^2), \tag{23}$$

$$x_{k,n}=t_n+\varepsilon_{k,n}t_n\ ;\ \varepsilon_{k,n}\sim\mathcal{N}(0,\sigma^2)\ \leftrightarrow\ x_{k,n}\sim\mathcal{N}(t_n,t_n\sigma^2). \tag{24}$$

That is, one of our sources is subject to additive Gaussian noise, one is subject to multiplicative Gaussian noise, and one is subject to additive heteroscedastic Gaussian noise with variance proportional to the mean. Again, we varied the noise standard deviation between $0 \leq \sigma \leq 0.5$.

The results from these two analyses are presented in Figure 7, again with 4% binning resolution. The linear TC analysis fails completely—the resulting estimators of total error variance and total correlation are essentially random and have no relationship with the true values of the statistics. Indeed, when we reran the linear analysis using several different truth and noise samples, the values of the estimators show no systematic relationship with the true statistics (not shown). In contrast, the estimates of the nonlinear total error and total information statistics are generally well behaved, even if the nonlinear estimators only bound the values of the true statistics due to the unconstrained nature of the problem. Of course, it is possible to know in advance that standard TC will fail here, simply by noting that the different measurement sources are not linearly related, however the point of this example is that the nonparametric approach can handle this type of situation.

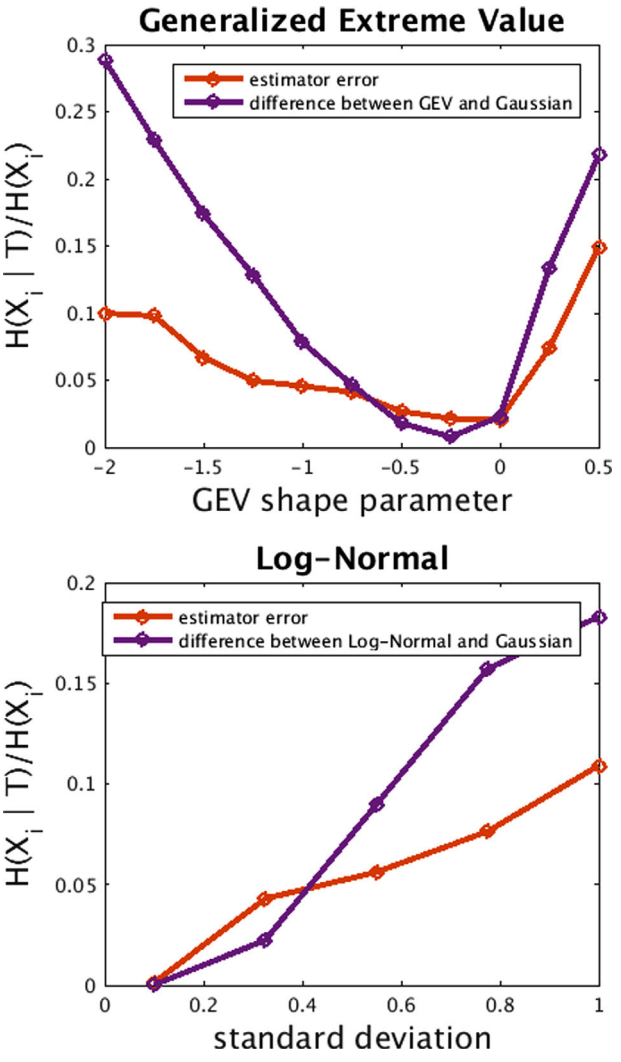

Figure 6. Comparison between underdetermination in nonparametric TC versus the ability to account for non-Gaussian errors. The orange line is the difference between the true and estimated total error terms, which is due to the inequality in equations (13)–(15): $\frac{H(X_i|X_j,X_k)}{H(X_i)}-\frac{H(X_i|T)}{H(X_i)}$. The purple line is the absolute difference between true total errors for non-Gaussian errors versus Gaussian errors with the same variances. All errors are additive.

3.2. Evaluating SMAP Retrievals Using Sparse In Situ Networks

Gridded soil moisture products from NASA's Soil Moisture Active Passive (SMAP) satellite mission [*Entekhabi et al.*, 2010] are validated against two primary types of in situ data. The first type of in situ data come from sparse validation networks, where there is typically only one in situ sensor per 36 km^2 SMAP pixel, and the second type of in situ data come from so-called core sites, where there are generally tens of in situ sensors per SMAP pixel and some confidence in the representativeness of those sensors for the larger domain. Sparse networks have the advantage of covering a larger number of locations on the Earth's surface, but are less reliable than the core sites because of

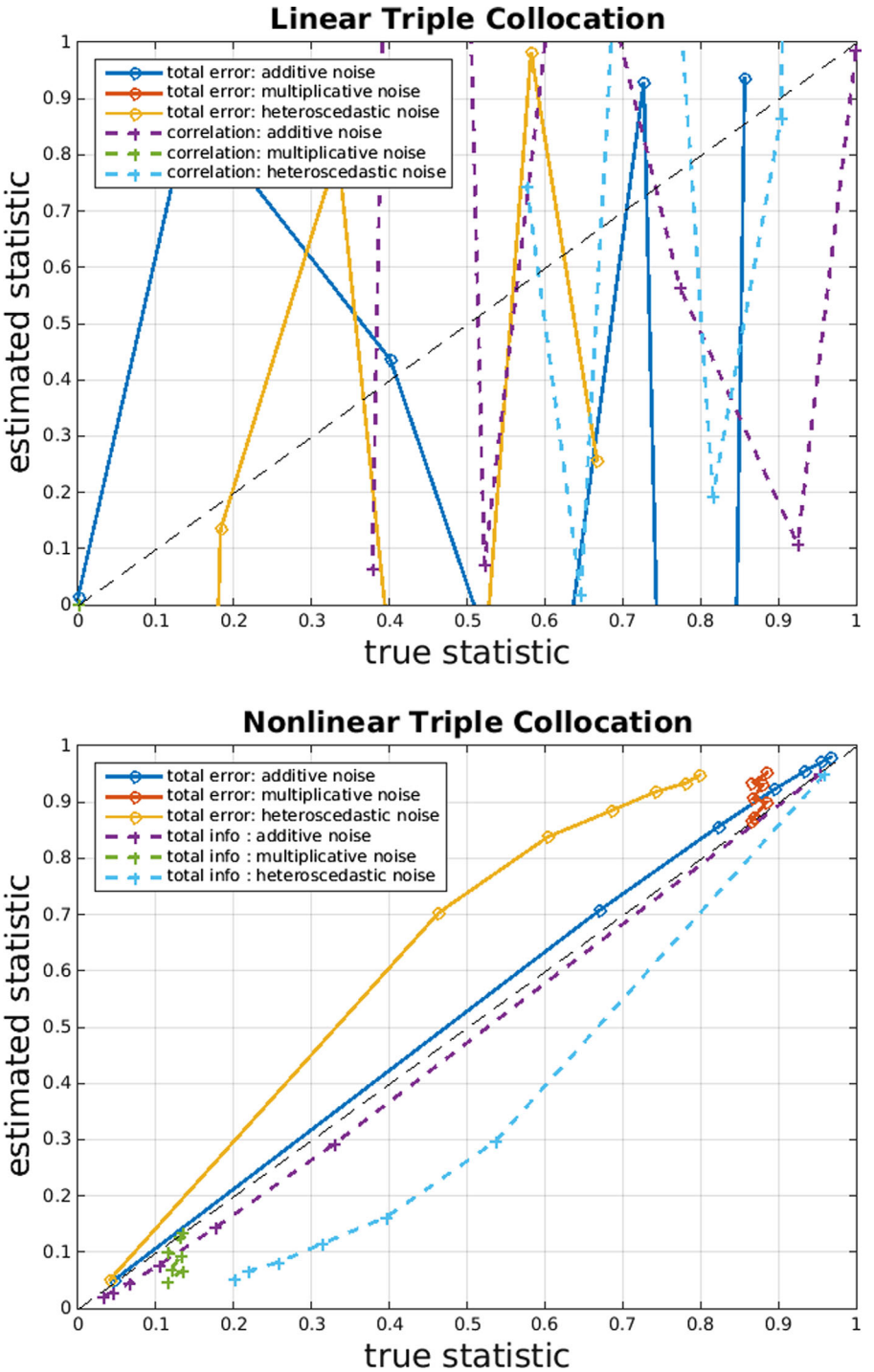

Figure 7. Results from the (top) linear and(bottom) nonlinear TC analyses where the three measurement sources are subject to noise that is: (i) additive Gaussian, (ii) multiplicative Gaussian, and (iii) heteroscedastic Gaussian.

the large difference in spatial support of individual soil moisture sensors (on the order of centimeters) and SMAP (on the order of kilometers), whereas soil moisture can vary significantly at scales on the order of meters.

Linear TC is one of the primary tools used for evaluating SMAP retrievals against the sparse network of in situ measurement locations that make their data available to the SMAP mission [e.g., *Scipal et al.*, 2010;

Chen et al., 2017]. To accomplish this, the three measurement sources are typically (i) SMAP retrievals, (ii) the in situ network, and (iii) a land surface hydrology model forced by observed precipitation.

The purpose of our final application example is to ask about the reliability of TC statistics to validate SMAP retrievals using a sparse network. To do this we will use data from five of the SMAP core sites, so that we may obtain from in situ data alone a relatively reasonable estimate of the spatially averaged soil moisture in an area that is at least the same order of magnitude as the SMAP pixel, which we will take as the "true" state of the system. We can then compare the actual linear and nonlinear triple colocation statistics calculated against this spatial average with statistics calculated using data from an individual in situ sensor, the SMAP retrievals, and outputs from a land surface hydrology model.

This type of experiment has been done before to check the validity of linear TC for evaluating SMAP using sparse network data [*Chen et al*., 2017]. We will reproduce that experiment, which uses model data from the European Center for Medium Range Weather Forecasts (ECMWF) and area-weighted averages of the in situ instruments as the "truth." Our questions here are (i) about the extent to which the linear approximations affect our ability to reliably estimate SMAP error and information content with TC against a sparse network, and (ii) whether we can improve on this analysis using the nonparametric technique outlined in this paper.

To preempt the results of this analysis, linear TC is relatively effective for this application. However, nonlinear TC produces total error and total information estimates that are between 1 and 2 orders of magnitude more accurate than their linear counterparts at a data resolution of 0.04 (m^3/m^3), which is the nominal (i.e., baseline mission requirement) SMAP accuracy.

To begin, we assessed whether there were enough data to apply nonparametric TC at a particular resolution. To do this, we partitioned the data into ten random samples of each of 10 different sample sizes at each core site. We did this at bin widths of 0.10, 0.04, and 0.02 (m^3/m^3). The objective is to see how many samples are required to achieve stable estimators at different data resolutions. In general, larger bin-widths require fewer data to achieve stable empirical density functions and stable integrations.

Results from this analysis are shown in Figure 8. What we see is that the estimators generally stabilize around the maximum sample size. Stability is not prefect, but the stable estimators (as a function of sample size) are achieved using about half of the available data at each site for bin-widths of about 0.02 (m^3/m^3) or greater. This means that we have sufficient data to estimate stable nonparametric statistics at the desired resolution of 0.04 (m^3/m^3) at all sites. Notice that this analysis of the tradeoff between data volume and resolution does not require any reference to truth, and therefore can be applied in any real-world situation.

Figure 9 compares the statistics of linear and nonlinear normalized total error and total correlation calculated assuming that the spatial average of all in situ measurements at each core site represents the true soil moisture versus the linear and nonlinear estimators of these statistics derived from the measurement sources (in situ, SMAP, and ECMWF) at the single in situ points. Results are presented separately for five of the SMAP core sites that are run by the USDA Agricultural Research Service. The total mean squared error for these estimators over all instruments at all sites are given in the first two columns of Table 1.

Because spatially averaged core site data are not a perfect representation of truth, the last two columns of Table 1 also report the mean squared errors for estimators of both linear and nonlinear total error and total correlation statistics calculated against (linear and nonlinear) TC statistics calculated using the spatial averages at each site as one of the measurements (instead of an individual in situ gage). That is, we estimated the TC statistics for ECMWF, SMAP, and a single in situ probe, and also the TC stats for ECMWF, SMAP, and the spatial average of in situ probes, and then compared these two sets of statistics. Discrepancies between these two sets of total error and total correlation statistics for both linear and nonlinear TC are in the last two columns of Table 1.

The main takeaway from Figure 9 is that the linear estimators do track the true linear statistics at three of the core sites (Little Washita, Fort Cobb, and Little River), but that at two of the core sites (Walnut Gulch and Reynolds Creek), they are unreliable. On the other hand, the nonlinear estimators are more-or-less well behaved across all sites. The main takeaway from Table 1 is that the total mean squared errors of the nonlinear estimators is almost consistently an order of magnitude lower than those of their linear counterparts when the spatial in situ average is taken as truth, and is almost 2 orders of magnitude lower when the spatial average is used as an independent measurement source. Again these results are for a data resolution of

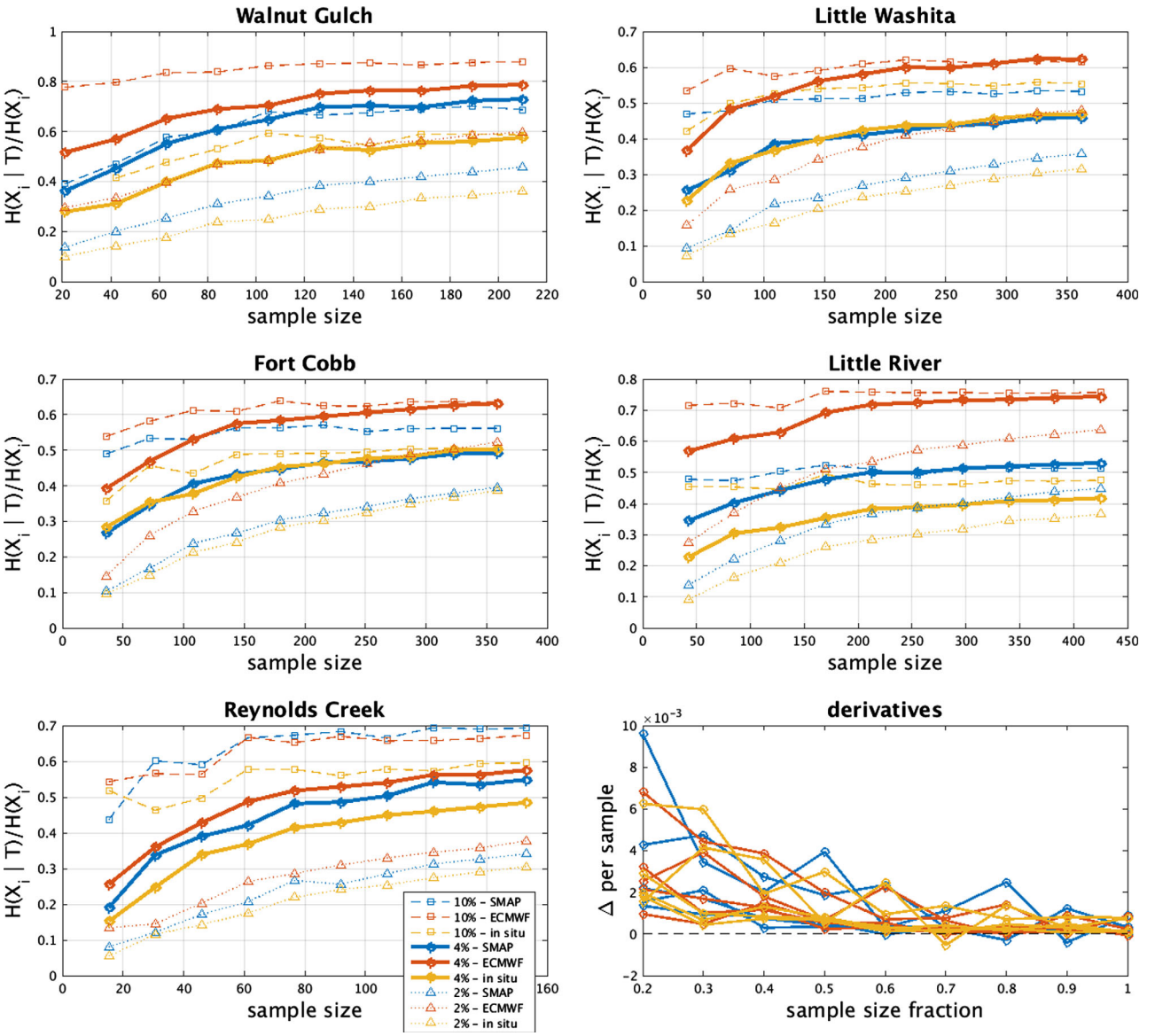

Figure 8. Assessment of the stability of total error estimates from nonparametric TC as a function of sample size and bin width. Plotted are mean estimators over ten random samples at each sample size. Bold lines represent a bin resolution of 0.04 (m^3/m^3), which is the baseline accuracy requirement for the SMAP mission, and the dotted lines represent bin resolutions of 0.02 (m^3/m^3) and 0.10 (m^3/m^3). The bottom-right subplot shows the (numerical) derivatives of statistic value as a function of sample size—derivatives at all sites are shown in this final subplot.

0.04 (m^3/m^3), however at a resolution of 0.02 (m^3/m^3) the mean squared errors of the nonlinear statistics are still everywhere lower than the means squared errors of the linear statistics by at least a factor of six (not shown).

One interesting feature of the results in Figure 9 is the comparison between linear and nonlinear statistics at Reynolds Creek. In recent TC soil moisture analyses [e.g., *Chen et al.*, 2017], Reynolds Creek statistics were dropped because it was reported that data from this site violate the independence assumptions. Here the conditional dependence of the three data sources was on average 0.16 (nats/nats) compared to an average of 0.42 (nats/nats) of total nonlinear correlation, which does mean that the independence assumption is violated. However, from Figure 9, we see that this violation of the independence assumption is not the primary cause of spurious (linear) total error and total correlation estimates at Reynolds creek—in fact, the problem is at least partially due to the nonlinearity of the measurement operators at this site.

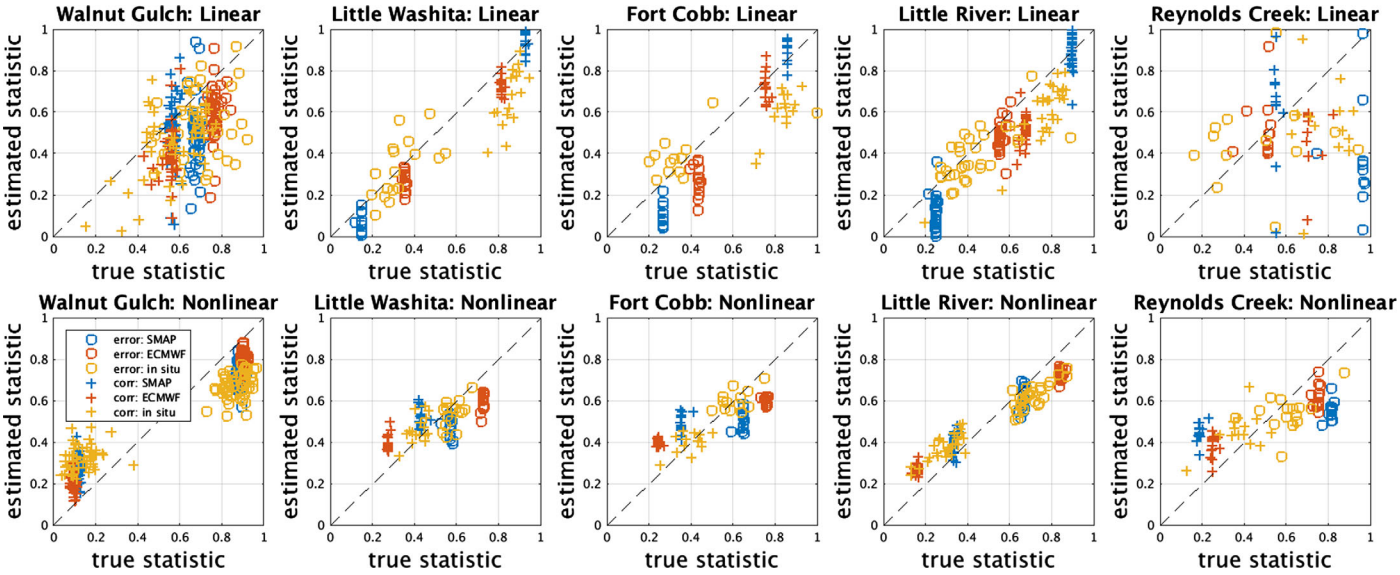

Figure 9. Scatterplots of true (linear and nonlinear) statistics of normalized total error and total correlation versus their sample-based estimators at five SMAP core validation sites. The true statistics were calculated against an area-weighted average of all in situ points, whereas the statistical estimators were each calculated for the respective measurement source at any single in situ point.

4. Summary and Discussion

One especially interesting aspect of this nonparametric triple collocation technique is the sense in which it alleviates the representativeness error problem discussed by *Gruber et al*. [2016]. In standard applications of TC, it is required that the three measurement sources observe essentially the same quantity, and so errors arise when, for example, soil moisture is observed at different scales (e.g., kilometer scales for models and remote sensing and centimeter scales for in situ probes). *Gruber et al*. [2016] gave a detailed discussion of this problem. Here we do not require that the three measurement sources observe the same variable, only whatever three phenomenon are observed have some probabilistic relationship with each other. That is, the measurement sources may be related by Markov chains of the form $T \rightarrow Y \rightarrow X_i$, $T \rightarrow Z \rightarrow X_j$, and $T \rightarrow W \rightarrow X_k$, where Y, Z, and W are entirely distinct system properties. The method we propose accounts for *any probabilistic relationship* between the three measurement sources.

It is important to understand that imposing strong parametric assumptions in our statistical methods and models often allows us to make strong claims within the context of those assumptions; however, such assumptions do not allow us to understand the full information content of our data. In the case of TC, assuming linear measurement models allows us to estimate *exactly* certain absolute error properties of our

Table 1. Mean Squared Errors in Statistics of Linear and Nonlinear Total Normalized Error and Total Correlation as Calculated Across All In Situ Instruments At All Five USDA SMAP Core Sites, With Area-Weighted Averages of the In Situ Instruments at Each Site Taken as "Truth"[a]

		Spatial Average as Truth		Spatial Average as a Measurement Source	
		Linear	Nonlinear	Linear	Nonlinear
SMAP	Error	0.419	0.026*	0.312	0.004*
	Corr	0.317	0.026*	0.312	0.004*
ECMWF	Error	0.296	0.016	0.278	0.003*
	Corr	0.291	0.016	0.278	0.003*
In situ	Error	0.126	0.022	0.132	0.017*
	Corr	0.049	0.022*	0.132	0.017*

[a]Asterisks indicate a statistically significant difference between linear and nonlinear estimators according to a one-sided *t*-test at α=0.05, p=0.01.

measurement sources, but this claim only holds within the context of these linearity assumptions, which will always be an approximation of any real, complex physical system.

Instead of looking for absolute answers constrained by unrealistic approximations, we advocate that a more relaxed question is "what is the best we can do with the available information?" Our purpose here is to demonstrate the answer to that question in the context of a three-source measurement problem. In support of this more general approach to understanding the information content and error properties of measurement data, we made three arguments:

1. Philosophically, we argued that information metrics are preferable to other types of performance metrics (e.g., variance-based) because they are more accurate descriptions of the epistemic situation that results from collecting observation data. This argument rests on the fact that information theory provides the only aptly quantitative description of the knowledge obtained from collecting measurements under probability theory.
2. Theoretically, we showed that the strong assumptions and approximations of TC (additivity and ignoring higher-order moments) are equivalent to a priori assignment of certain properties of the information content of the measurements, which again in our opinion, is largely unhelpful if our objective is to understand what we can learn from data.
3. Empirically, we showed that strong and unrealistic constraints are probably not necessary for many TC problems. In most of our test cases, the nonparametric estimators generally outperformed their linear counterparts in terms of estimating the true quantities of interest.

The major tradeoff is that the nonlinear statistics require us to specify a precision for our measurement data. However, this effect can be assessed relative to the sample size required to achieve stable empirical estimators.

Notice that the first two arguments are interactive, however the empirical argument holds whether the scientist does or does not value a rigorous epistemological basis for their performance metrics. We therefore imagine that most scientists would be interested in the empirical demonstrations because these show examples of how nonparametric TC is a practical methodology that is at least potentially more reliable than linear TC for at least certain applications. That being said, it is our opinion that the philosophical argument is the most valuable contribution of this paper, as this piece of epistemology can be generalized to a wide array of statistical methods, models, and applications [e.g., *Nearing et al*., 2013; *Nearing and Gupta*, 2015].

Acknowledgement
Funding was provided by the NASA Terrestrial Hydrology Program. All data and code are publically available on GihHub at https://github.com/greyNearing/triple_collocation.

References

Caires, S., and A. Sterl (2003), Validation of ocean wind and wave data using triple collocation, *J. Geophys. Res.*, *108*(C3), 3098, doi:10.1029/2002JC001491.

Chen, F., W. T. Crow, A. Colliander, M. H. Cosh, T. J. Jackson, R. Bindlish, R. H. Reichle, S. K. Chan, D. D. Bosch, and P. J. Starks (2017) Application of triple collocation in ground-based validation of Soil Moisture Active/Passive (SMAP) level 2 data products, *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, *10*(2), 489–502.

Cox, R. T. (1946), Probability, frequency and reasonable expectation, *Am. J. Phys.*, *14*, 1–13.

Entekhabi, D., et al. (2010) 'The Soil Moisture Active Passive (SMAP) Mission', *Proc. IEEE*, *98*(5), 704–716, doi:10.1109/JPROC.2010.2043918.

Gruber, A., C.-H. Su, S. Zwieback, W. Crow, W. Dorigo, and W. Wagner (2016), Recent advances in (soil moisture) triple collocation analysis, *Int. J. Appl. Earth Observ. Geoinform.*, *45*, 200–211.

Knuth, K. H. (2005), Lattice duality: The origin of probability and entropy, *Neurocomputing*, *67*, 245–274.

McColl, K. A., J. Vogelzang, A. G. Konings, D. Entekhabi, M. Piles, and A. Stoffelen (2014), Extended triple collocation: Estimating errors and correlation coefficients with respect to an unknown target, *Geophys. Res. Lett.*, *41*, 6229–6236, doi:10.1002/2014GL061322.

Miyaoka, K., A. Gruber, F. Ticconi, S. Hahn, W. Wagner, J. Figa-Saldaña, and C. Anderson (2017), Triple collocation analysis of soil moisture from Metop-A ASCAT and SMOS against JRA-55 and ERA-Interim, *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, *10*(5), 2274–2284.

Nearing, G. S., and H. V. Gupta (2015), The quantity and quality of information in hydrologic models, *Water Resour. Res.*, *51*, 524–538, doi:10.1002/2014WR015895.

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

Paninski, L. (2003), Estimation of entropy and mutual information, *Neural Comput.*, *15*, 1191–1253.

Roebeling, R., E. Wolters, J. Meirink, and H. Leijnse (2012), Triple collocation of summer precipitation retrievals from SEVIRI over Europe with gridded rain gauge and weather radar data, *J. Hydrometeorol.*, *13*(5), 1552–1566.

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

Scipal, K., W. Dorigo, and R. de Jeu (2010), Triple collocation: A new tool to determine the error structure of global soil moisture products, in *2010 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, Honolulu, Hawaii.

Shannon, C. E. (1948), A mathematical theory of communication, *Bell Syst. Tech. J.*, *27*(3), 379–423.

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

Vogelzang, J., and A. Stoffelen (2012), Triple collocation, *EUMETSAT Rep. TR KN, 21*, pp. v1. [Available at http://research.metoffice.gov.uk/research/interproj/nwpsaf/scatterometer/TripleCollocationNWPSAF.]

Yilmaz, M. T., and W. T. Crow (2014), Evaluation of assumptions in soil moisture triple collocation analysis, *J. Hydrometeorol.*, *15*(3), 1293–1302.