Remote Sensing of Environment 176 (2016) 31–42

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Triple collocation for binary and categorical variables: Application to validating landscape freeze/thaw retrievals

Kaighin A. McColl [a,*], Alexandre Roy [c], Chris Derksen [d], Alexandra G. Konings [a], Seyed Hamed Alemohammed [a], Dara Entekhabi [a,b]

[a] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[b] Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[c] Centre d'Application et de Recherches en Télédétection, Université de Sherbrooke, Sherbrooke, QC J1K 2R1, Canada
[d] Environment Canada, Toronto, ON M3H 5T4, Canada

ARTICLE INFO

Article history:
Received 3 September 2015
Received in revised form 11 January 2016
Accepted 16 January 2016
Available online 23 January 2016

Keywords:
Triple collocation
Freeze/thaw classification
SMAP
Aquarius

ABSTRACT

Triple collocation (TC) can be used to validate observations of a continuous geophysical target variable when the error-free true value is not known. However, as we show in this study, naïve application of TC to categorical target variables results in biased error estimates. The bias occurs because the categorical variable is usually bounded, introducing correlations between the errors and the truth, violating TC's assumptions. We introduce Categorical Triple Collocation (CTC), a variant of TC that relaxes these assumptions and may be applied to categorical target variables. The method estimates the rankings of the three measurement systems for each category with respect to their balanced accuracies (a binary-variable performance metric). As an example application, we estimate performance rankings of landscape freeze/thaw (FT) observations derived from model soil temperatures, in-situ station air temperatures and satellite-observed microwave brightness temperatures in Alberta and Saskatchewan, Canada. While rankings vary spatially, in most locations the model-based FT product is ranked the highest, followed by the satellite product and the in-situ air temperature product. These rankings are likely due to a combination of differences in measurement errors between FT products, and differences in scale. They illustrate the value in using a suite of different measurements as part of satellite FT validation, rather than simply treating in-situ measurements as an error-free 'truth'.

1. Introduction

Categorical variables belong to one of a set of exhaustive, mutually-exclusive categories, which may be ordered (in which case, the categorical variable is 'ordinal') or unordered ('nominal'). For many geophysical variables, it is convenient to consider the variable to be categorical rather than continuous. Examples include land cover type (Friedl et al., 2002), cloud presence/absence (Ackerman et al., 1998), wildfire burned area status (Roy, Boschetti, Justice, & Ju, 2008) and landslide occurrence (Metternicht, Hurni, & Gogu, 2005). Models, satellites and in-situ observations (or "measurement systems") are used to monitor and understand these variables, but each system contains its own errors. A common question is: which system has the best performance ranking with respect to an appropriate validation metric (Entekhabi, Reichle, Koster, & Crow, 2010)?

One measurement system is usually assumed a priori to be the error-free "truth" system, with other systems judged in comparison. However, the presence of inevitable errors in the "truth" system, along with differences in support scale between systems, often make the performance rankings dependent on the choice of the "truth" system, an unsatisfactory outcome. Triple collocation (TC) is a technique for estimating the root-mean-squared-errors (Stoffelen, 1998) and correlation coefficients (McColl, Vogelzang, et al., 2014) of three measurement systems with respect to the unknown true value of a continuous target variable, without unrealistically treating any one system as error-free. It has been used for estimating errors in measurements of a wide range of continuous geophysical target variables, including sea surface temperature (e.g., O'Carroll, Eyre, & Saunders, 2008), wind speed and stress (e.g., Vogelzang, Stoffelen, Verhoef, & Figa-Saldaña, 2011), wave height (e.g., Janssen, Abdalla, Hersbach, & Bidlot, 2007), precipitation (Alemohammad, McColl, Konings, Entekhabi, & Stoffelen, 2015; Roebeling, Wolters, Meirink, & Leijnse, 2012), fraction of absorbed photosynthetically active radiation (D'Odorico et al., 2014), leaf area index (Fang, Wei, Jiang, & Scipal, 2012) and soil moisture (e.g., Draper et al., 2013; Miralles, Crow, & Cosh, 2010).

Applying triple collocation to categorical target variables, however, poses unique challenges. Problems arise because categorical variables are usually unordered and bounded. As we show in Section 2, these differences mean that key assumptions in TC are violated, biasing TC

* Corresponding author.
E-mail address: kmccoll@mit.edu (K.A. McColl).

http://dx.doi.org/10.1016/j.rse.2016.01.010
0034-4257/

error estimates. In Section 3, we describe a new approach – extending the work of Parisi, Strino, Nadler, and Kluger (2014) – called Categorical Triple Collocation (CTC) that relaxes the violated assumptions and provides performance rankings for measurements of categorical variables. In Sections 4 and 5, we demonstrate its utility by applying it to the problem of ranking the performances of model, in-situ and satellite estimates of landscape freeze/thaw (FT) state.

2. Deficiencies of classical TC

Triple collocation is a commonly used technique for estimating the mean-squared error MSE (Stoffelen, 1998) and correlation coefficient r (McColl, Vogelzang, et al., 2014) of a measurement or model estimate with respect to the unknown true value of the target variable. It requires observations of the target variable from three collocated measurement systems that are linearly related to the target variable. The error model is given by

$$X_i = \alpha_i + \beta_i T + \varepsilon_i \tag{1}$$

where X_i (for $i = 1,2,3$) are the observed measurements from the noisy measurement systems, T is the unknown true value of the target variable, ε_i is a zero-mean random error term and α_i, β_i are calibration parameters. X_i, ε_i and T are all random variables. It is further assumed that $\mathrm{Var}(\varepsilon_i)$ and $\mathrm{Var}(T)$ are fixed and do not vary in time. The same assumption is not strictly required for $\mathrm{E}(T)$, although many TC studies use climatological anomalies so that $\mathrm{E}(T)$ is approximately stationary. The three measurement systems used in the analysis could be, for example, a satellite retrieval, a model estimate and an in-situ observation of the target variable. To apply triple collocation, two additional assumptions must be satisfied:

(R1) the random errors between different measurement systems must be uncorrelated with each other (i.e., $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0, i \neq j$).
(R2) the random errors must not be state-dependent and must be uncorrelated with the target variable (i.e., $\mathrm{Cov}(\varepsilon_i, T) = 0$).

Classical TC suffers from several deficiencies when applied to categorical variables, arising from the facts that they may be unordered and/or strongly bounded. First, the additive, zero-mean error model implicitly imposes an ordering and is inappropriate for nominal (i.e., unordered) categorical variables. Second, even if we only consider ordinal (i.e., ordered) variables, the distribution of ε_i must depend on T to ensure that X_i does not take on values outside the bounded domain. This dependence violates (R2) and becomes more significant as the number of possible values the categorical variable may take on (i.e., the size of its support) decreases. Consider the case of binary variables, which only have two elements in their support (i.e., $X_i, T \in \{-1,1\}$). As shown in Appendix A, this limited support induces non-negligible correlations between the errors and target variable such that (R1) and (R2) are always strongly violated for the binary case. In particular, defining P_i to be the probability of an error occurring in measurement system i, we have

$$\mathrm{Cov}(\varepsilon_i, T) = -2P_i \tag{2}$$

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 4P_iP_j\mathrm{Var}(T) \tag{3}$$

which are non-zero for all non-trivial cases where $P_i > 0, P_j > 0$ and $\mathrm{Var}(T) > 0$. The observation that $\mathrm{Cov}(\varepsilon_i, T) < 0$ for categorical data has been widely noted in the econometrics literature in terms of 'mean reversion': errors tend to be biased towards the mean (e.g., Kapteyn & Ypma, 2007). The correlation between the errors and the truth then induces correlations between errors in the different measurement systems. These violations of (R1) and (R2) result in biased triple collocation error estimates.

3. Triple collocation for categorical variables

The flaws in classical TC when applied to categorical variables motivate the development of a new approach that uses an error model appropriate for unordered variables, and allows the errors and truth to be correlated. In this section, we will introduce a variant of TC for categorical variables that estimates performance rankings of three measurement systems with respect to a binary validation metric, the "balanced accuracy"

$$\pi = \frac{1}{2}(\psi + \eta) \tag{4}$$

where ψ is the measurement system sensitivity (i.e., the probability of the measurement being correct when the truth $T = 1$) and η is the measurement system specificity (i.e., the probability of the measurement being correct when $T = -1$). Unlike the simple accuracy metric μ (i.e., the probability of the measurement being correct over all cases), π avoids overestimating the quality of performance of biased classifiers on imbalanced datasets ($\mathrm{E}(T) \neq 0$), while still reducing to μ for balanced datasets. For example, consider a binary classifier which is biased, in that it always returns a classification of 1. If T is almost always 1, the biased classifier may still receive a high simple accuracy, even though it has no real predictive skill. In contrast, the balanced accuracy will more heavily penalize the classifier for the rare occurrences where $T = -1$ and the classification is incorrect. It is impossible to derive the actual balanced accuracy for each measurement system but, as will be shown, our approach allows calculation of a quantity that is proportional to the balanced accuracy for each measurement system. The relative sizes of this quantity between the three measurement systems can be used to determine relative performance rankings.

To handle unordered variables, instead of the linear regression framework adopted in classical TC, we use a classification framework. For each measurement system i and category k, define a binary classifier

$$X_i^k\left(T^k\right) = T^k + \varepsilon_i^k \tag{5}$$

where

$$T^k = \begin{cases} 1, & \text{if the true value belongs to class } k \\ -1, & \text{otherwise} \end{cases} \tag{6}$$

and

$$X_i^k = \begin{cases} 1, & \text{if the measured value belongs to class } k \\ -1, & \text{otherwise} \end{cases} \tag{7}$$

with $\varepsilon_i^k \in \{-2,0,2\}$, and dependent on the value of T^k to ensure that X_i^k does not take on a value outside the set $\{-1,1\}$. We may then assess the performance of the measurement system separately for each category. For instance, say we are validating a landcover type categorical variable, with the categories 'grassland', 'forest', 'desert' and 'other'. We can treat this as four different binary classification problems: 'grassland' vs 'not grassland', 'forest' vs 'not forest', 'desert' vs 'not desert' and 'other' vs 'not other'. This will result in four separate rankings for the four different categories. As a consequence, for example, the measurement system that is ranked the highest for 'grassland' may be ranked the lowest for 'desert'. There is no single, obvious way to combine these different rankings into a single ranking across categories. This is a general problem common to all categorical classification techniques. Hence, the problem of validating general categorical variables reduces to that of validating binary variables; we now drop the k superscript in our notation for convenience.

To allow the errors and truth to be correlated, we must relax the assumption in our previous error model (1) that $E(\varepsilon_i) = 0$, since this will now also depend on T. As shown in Appendix B, in general,

$$E(\varepsilon_i) = 2(1-\eta_i)p_T(T=-1) - 2(1-\psi_i)p_T(T=1) \quad (8)$$

where $\psi_i = \Pr(X_i = T|T=1)$ is the sensitivity and $\eta_i = \Pr(X_i = T|T=-1)$ is the specificity.

Another important difference with the classical TC error model is that $\mathrm{Var}(T)$ cannot be assumed to be fixed in time and independent of $E(T)$. Indeed, $\mathrm{Var}(T)$ is directly determined by $E(T)$, and will vary in time if $E(T)$ varies in time: for a time-varying binary target variable, $E(T|t) = 2p(t) - 1$ and $\mathrm{Var}(T|t) = 4p(t)(1-p(t))$ where $p(t) \equiv p_T(T=1|t)$ and t is time.

Building on a recent method (Parisi et al., 2014, hereafter P14), we propose a variant of TC (Categorical Triple Collocation, or CTC) that is valid for this model and, crucially, does not violate its assumptions when applied to categorical variables. The assumptions (R1) and (R2) are replaced with the following assumption:

(R1*) the random errors between different measurement systems must be conditionally independent ($\Pr(\varepsilon_i, \varepsilon_j|T) = \Pr(\varepsilon_i|T)\Pr(\varepsilon_j|T)$, for all $i \neq j$).

This is a less restrictive assumption than those required in classical TC, since the errors may now be dependent on T.

3.1. Stationary case (Parisi et al., 2014)

First, we review the approach given in P14, which is valid for stationary variables. In their derivation, apart from (R1*), they further require stationarity: that is, $p(t)$, $E(T)$ and $\mathrm{Var}(T)$ are all constant in time (Fig. 1c). From these assumptions, they show that

$$Q_{ij} \equiv \mathrm{Cov}(X_i, X_j) = \begin{cases} 1 - E(X_i)^2, & \text{for } i = j \\ \mathrm{Var}(T)(2\pi_i - 1)(2\pi_j - 1), & \text{otherwise} \end{cases} \quad (9)$$

(equivalent to P14's Eq. (5)), where π_i is the balanced accuracy of the ith system. P14 consider a general problem involving M classifiers. For the special case relevant to triple collocation ($M = 3$), we have four unknowns (π_1, π_2, π_3 and $E(p(t))$; note that $E(X_i)$ can be estimated from the data), and six unique terms in the 3×3 covariance matrix estimate ($Q_{11}, Q_{12}, Q_{13}, Q_{22}, Q_{23}, Q_{33}$). However, only three of these terms (Q_{12}, Q_{13}, Q_{23}) can be expressed as equations in terms of the unknowns. Therefore, we have three equations and four unknowns, the system is underdetermined and there is no unique solution. However, if we define the change of variable $v_i = (2\pi_i - 1)\sqrt{\mathrm{Var}(T)}$, we can rewrite the system as

$$Q_{ij} \equiv \mathrm{Cov}(X_i, X_j) = \begin{cases} 1 - E(X_i)^2, & \text{for } i = j \\ v_i v_j, & \text{otherwise} \end{cases} \quad (10)$$

There are now three equations and three unknowns (v_1, v_2, v_3), so $\boldsymbol{v}$ is exactly determined, and can be shown to be

$$\boldsymbol{v} = \begin{bmatrix} \sqrt{\frac{Q_{12}Q_{13}}{Q_{23}}} \\ \sqrt{\frac{Q_{12}Q_{23}}{Q_{13}}} \\ \sqrt{\frac{Q_{23}Q_{13}}{Q_{12}}} \end{bmatrix} \quad (11)$$

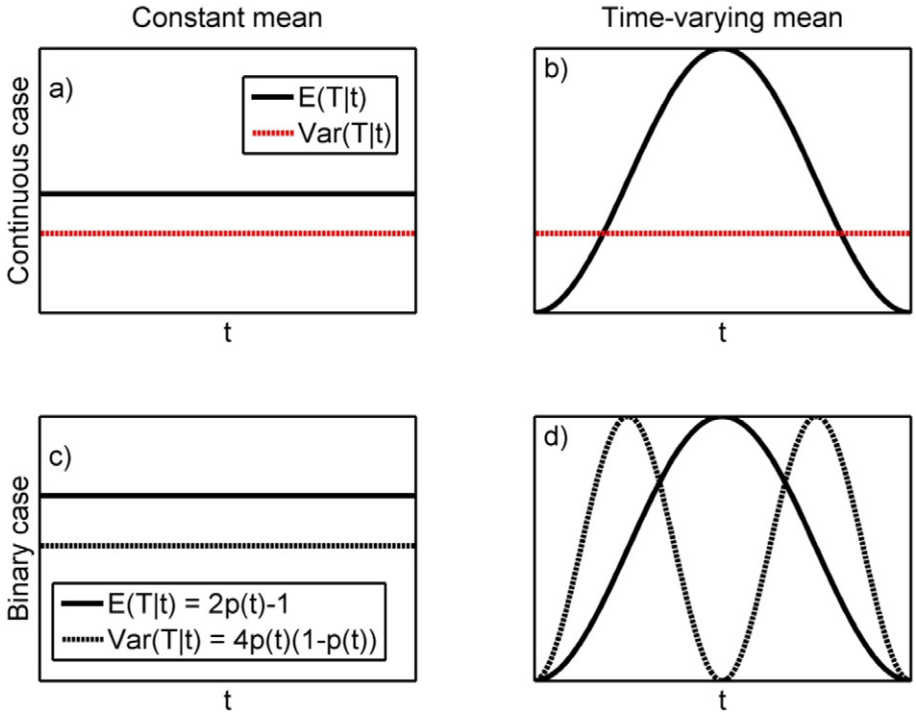

Fig. 1. Schematic of the different models of the truth $T(t)$ used in classical TC, P14, and CTC, where t is time. In classical TC, the variables are continuous (top row). The true variance $\mathrm{Var}(T|t)$ is assumed constant and independent of the true mean $E(T|t)$, which may be constant (a), e.g., when TC is applied to climatological anomalies. It may also vary in time (b), e.g., when TC is applied directly to time series. When $T(t)$ is a binary variable (bottom row), $\mathrm{Var}(T|t)$ is directly determined by $E(T|t)$ (this dependence is represented in the schematic by coloring both $E(T|t)$ and $\mathrm{Var}(T|t)$ black). The dependence arises because they are both functions of $p(t) = \Pr(T = 1|t)$. P14 provide a TC-variant for the case where $p(t)$ is constant in time (c). However, in geophysical applications, $p(t)$ often varies in time, e.g., seasonally (d), implying that $E(T|t)$ and $\mathrm{Var}(T|t)$ also vary seasonally. CTC is a generalization of the approach of P14 to the binary, time-varying case (d).

Since v_i is a monotonic increasing function of π_i, sorting $\boldsymbol{v}$ yields the measurement system rankings in terms of balanced accuracy (except in the degenerate cases where $p(t) = 0$ or $p(t) = 1$ for all t; or $\pi_i = 0.5$ for any i). For instance, if we have $v_2 > v_1 > v_3$, then measurement system 2 displays the best performance, followed by system 1 and system 3, which displays the poorest performance. The equations make intuitive sense by rewarding system i when it covaries strongly with systems j and k, and rewarding it even more so when systems j and k do not covary with each other much. By replacing (R1) and (R2) with the less restrictive (R1*), however, we pay the price of only obtaining rankings of the measurement systems in terms of a validation metric, rather than absolute values of the metric itself. Our simpler derivation of $\boldsymbol{v}$ is fully equivalent to that given in P14 (where $\boldsymbol{v}$ is defined as the eigenvector of a rank-one matrix $\boldsymbol{R}$ with identical off-diagonal elements to those of $\boldsymbol{Q}$). The simpler exposition is possible for the $M = 3$ case that we consider because the number of unknowns ($M = 3$) equals the number of equations available (i.e., half the number of off-diagonal terms in the covariance matrix, $M(M\text{-}1)/2 = 3$), resulting in an exact solution (Eq. (11)) not possible for $M > 3$, where the problem is overdetermined.

3.2. Non-stationary case (CTC)

In geophysical applications, the samples are often time series that contain significant seasonal variation, and are consequently not identically-distributed: that is, $p(t)$, $\mathrm{E}(T|t)$ and $\mathrm{Var}(T|t)$ all vary significantly in time (Fig. 1d). This means the relationship between the observed covariances and balanced accuracies derived in P14 (Eq. (9)) is now changing in time, appearing to pose a problem for obtaining a single set of rankings across time. In Appendix C, we relax the assumption of stationarity made in P14 to derive a more general version of Eq. (9):

$$Q_{ij} \equiv \mathrm{Cov}(X_i, X_j) = \begin{cases} 1 - \mathrm{E}(\mathrm{E}(X_i|t))^2, & \text{for } i = j \\ 4\mathrm{E}(p(t))(1 - \mathrm{E}(p(t)))(2\pi_i - 1)(2\pi_j - 1), & \text{otherwise} \end{cases} \quad (12)$$

As expected, Eq. (12) reduces to Eq. (9) when $p(t)$ is constant in time. Remarkably, relaxing the assumption of stationarity used in P14 only affects the constant of proportionality between the covariances and the balanced accuracies. Following a similar approach as before, if we define the (different) change of variable $w_i = 2(2\pi_i - 1)\sqrt{\mathrm{E}(p(t))(1 - \mathrm{E}(p(t)))}$, we can rewrite the system as

$$Q_{ij} \equiv \mathrm{Cov}(X_i, X_j) = \begin{cases} 1 - \mathrm{E}(\mathrm{E}(X_i|t))^2, & \text{for } i = j \\ w_i w_j, & \text{otherwise} \end{cases} \quad (13)$$

which can be solved for $\boldsymbol{w}$, to obtain

$$\boldsymbol{w} = \begin{bmatrix} \sqrt{\frac{Q_{12}Q_{13}}{Q_{23}}} \\ \sqrt{\frac{Q_{12}Q_{23}}{Q_{13}}} \\ \sqrt{\frac{Q_{23}Q_{13}}{Q_{12}}} \end{bmatrix} = \boldsymbol{v} \quad (14)$$

We have identified a solution for the non-stationary case, and this solution turns out to be identical to the solution for the stationary case! This is a surprising result given our initial expectations that non-stationarity should pose problems for obtaining a solution. Hence, we have generalized this approach to a wider class of problems compared to those shown in P14. In particular, we have shown that it can be applied to non-stationary geophysical variables and can be considered a form of triple collocation for categorical variables. Fig. 1 summarizes some of the similarities and differences between CTC, P14 and classical TC.

CTC can be summarized in three steps:

1. Calculate the sample 3×3 covariance matrix $\boldsymbol{Q}$ from the observations $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3$.
2. Use $\boldsymbol{Q}$ and Eq. (14) to estimate $\boldsymbol{w}$.
3. Sort $\boldsymbol{w}$ (in descending order) to obtain rankings.

3.3. Simulation studies

A synthetic example is used to illustrate the application of CTC (Fig. 2). In this example, the geophysical binary variable follows an annual cycle similar to that of landscape freeze/thaw state, the variable of interest in the next section. Very early and very late in the year (i.e., during the boreal winter), $p(t) \sim 1$ and it is therefore highly likely that $T = 1$ (i.e., the landscape is frozen) during these periods. During the middle of the year (i.e., the boreal summer), $p(t) \sim 0$ and it is therefore highly likely that $T = -1$ (i.e., thawed) during this time. As $p(t)$ transitions between these two deterministic end-points, T shows greater random variability, peaking at $p(t) = 0.5$ (around boreal spring and fall). For the synthetic example in Fig. 2, a single, 52-week realization of the seasonally-varying binary random-variable is used as the synthetic truth T. Observations from three measurement systems (for instance, a satellite, a model, and an in-situ station) with conditionally-independent errors are simulated by inserting errors in the synthetic truth series. More specifically, for measurement system 1, errors are randomly inserted such that the sensitivity $\psi_1 = \Pr(X_1 = T|T = 1) = 0.8$, the specificity $\eta_1 = \Pr(X_1 = T|T = -1) = 0.6$ and the

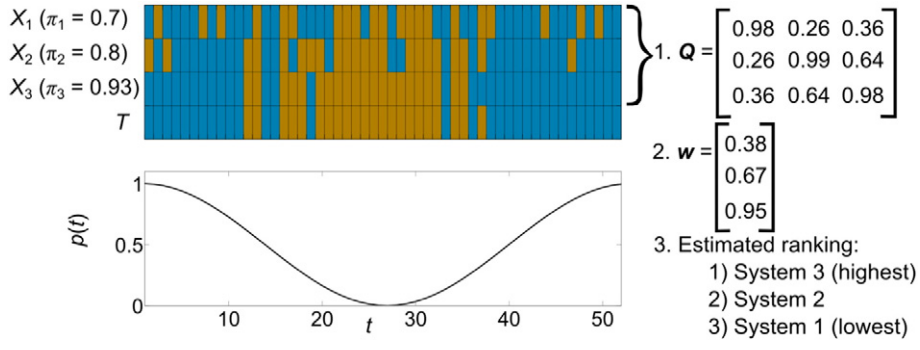

Fig. 2. Application of CTC to a synthetic example. In this example, $p(t)$ follows a seasonal cycle over one year (bottom left), where t is time (weeks). A single, 52-week realization of the seasonally-varying binary random-variable is used as the synthetic truth T (top left grid: bottom row, where values of 1 and −1 are colored blue and light brown, respectively). Synthetic observations from three different measurement systems ($\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$) with prescribed balanced accuracies ($\pi_1 = 0.7$, $\pi_2 = 0.8$, $\pi_3 = 0.93$) are generated by inserting errors into the truth series (top left grid: top three rows). The three-step CTC procedure is followed to estimate rankings for the three systems with respect to their balanced accuracies from the synthetic data: 1) The sample covariance matrix $\boldsymbol{Q}$ is estimated from $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$ (top right) 2) Eq. (14) is used to estimate $\boldsymbol{w}$ (middle right) 3) the order of terms in $\boldsymbol{w}$ is used to determine the rankings (bottom right). The estimated CTC ranking correctly reflects the (unknown) differences in balanced accuracy.

balanced accuracy is therefore $\pi_1 = \frac{1}{2}(\psi_1 + \eta_1) = 0.7$ (by definition in Eq. (4)). For measurement system 2, $\psi_2 = 0.9$, $\eta_2 = 0.7$ and therefore $\pi_2 = 0.8$; and for measurement system 3, $\psi_3 = 0.98$, $\eta_3 = 0.88$ and therefore $\pi_3 = 0.93$. Therefore, in this synthetic example, system 3 displays the best performance with respect to the balanced accuracy metric, followed by system 2, with system 1 displaying the poorest performance. This is evident in Fig. 2, where $\boldsymbol{X_3}$ is almost identical to $\boldsymbol{T}$, in contrast to $\boldsymbol{X_1}$, which is substantially different to $\boldsymbol{T}$. CTC is able to identify the correct performance ranking of the three measurement systems in this case, without access to the truth $\boldsymbol{T}$ (Fig. 2).

CTC relies on an estimate of the covariance matrix $\boldsymbol{Q}$ from a finite sample. For small sample sizes, the sample estimate of $\boldsymbol{Q}$ may be noisy and result in incorrect CTC performance rankings. Furthermore, if measurement systems with very low skill are used in a CTC analysis, some elements of $\boldsymbol{Q}$ may be very small and thus more susceptible to noise (for the limiting case with $\pi_1 = \pi_2 = \pi_3 = 0.5$, the off-diagonal terms of $\boldsymbol{Q}$ in Eq. (12) are all zero and CTC is degenerate). To gain some understanding of these effects on CTC performance rankings, sensitivity tests are performed using the example presented in Fig. 2 as a reference case (Fig. 3). In the first scenario, the balanced accuracy of the highest-performing measurement system (system 3) is systematically increased from its reference value ($\pi_3 = 0.93$), and the sample size is increased from one to ten years (with the same seasonally-varying $p(t)$ relationship used for each subsequent year as that used in Fig. 2). For each case, 500 replicates of the truth T are sampled. Each replicate is used to generate synthetic observations from three measurement systems, with randomly-assigned errors but fixed balanced accuracies. CTC is used to estimate the rankings of the three systems, as in the reference case, for each replicate. Due to sampling errors, for some replicates, CTC does not identify the correct ranking. We summarize the frequency with which this occurs by calculating, for each case, the proportion of replicates in which CTC correctly identifies the highest-ranked measurement system (Fig. 3, left). In all cases, in this scenario, CTC does much better at identifying the highest-ranked measurement system than random guessing (corresponding to a probability of 0.33, the lowest value in the colorbar in Fig. 3). As expected, CTC sampling error decreases as sample sizes increase, with CTC almost always correctly identifying the highest-ranked measurement system when ten years of weekly observations are used. CTC is also less vulnerable to sampling error as π_3 increases. This makes intuitive sense: it is easier to identify the best system when it is much better than its competitors, and more difficult when they are all similar.

In the second scenario (Fig. 3, right), the lowest-ranked system in the reference case (system 1) is changed so that $\pi_1 = 0.51$. This simulates the inclusion of a very poor measurement system in a CTC analysis, with predictive skill only marginally better than random guessing. In this scenario, the CTC solution is very close to degenerate, and the small elements of $\boldsymbol{Q}$ are easily overwhelmed by even small levels of noise due to sampling error. As such, CTC is unable to perform better than random guessing for this near-degenerate case, even with large sample sizes or large values for π_3. Therefore, as expected, using low-skill measurement systems as inputs will result in low-accuracy performance rankings from CTC.

4. Example application: validating landscape freeze/thaw state

We apply CTC to the validation of an important categorical geophysical variable: landscape freeze/thaw (FT) state. The landscape FT state across boreal regions has important regional and global environmental implications. It modulates the partitioning of land-surface energy fluxes and, consequently, the state of the atmospheric column and regional weather patterns (Betts, Viterbo, Beljaars, & van den Hurk, 2001). Globally, the timing of FT transitions bounds the growing season for vegetation, with important implications for exchanges of carbon between the land and atmosphere in a changing climate (Goulden et al., 1998).

Monitoring landscape FT state globally is difficult because it requires knowledge of the state of soil, snow and vegetation over a vast region, including remote areas that are difficult to access. Low-frequency radar and radiometer observations from satellites offer a promising means for monitoring landscape FT globally, since the FT state has a strong impact on water permittivity and hence microwave scattering and emission from the soil and vegetation (Way et al., 1990; Wegmüller, 1990). Satellite measurements at L-band from the NASA Soil Moisture Active Passive (SMAP; Dunbar et al., 2015; Entekhabi, Njoku, et al., 2010) and ESA Soil Moisture and Ocean Salinity (SMOS; Rautiainen et al., in press) mission provide regular retrievals of soil moisture and FT state. It is necessary to characterize FT product errors for satellite validation purposes, and so that the FT observations may be assimilated into land-surface models (Farhadi, Reichle, De Lannoy, & Kimball, 2014).

Binary retrievals of FT state from satellite measurements are difficult to validate, because unlike variables such as soil moisture, 'freeze' and 'thaw' states are not measured directly in-situ; rather, they must be inferred from in-situ observations of air and/or soil temperatures. The influence of vegetation and snow cover (particularly wet snow) on radar and radiometer measurements must also be considered. In previous studies, satellite FT estimates were generally validated by comparison with meteorological station observations or model outputs of continuous proxies such as air temperature or soil temperature

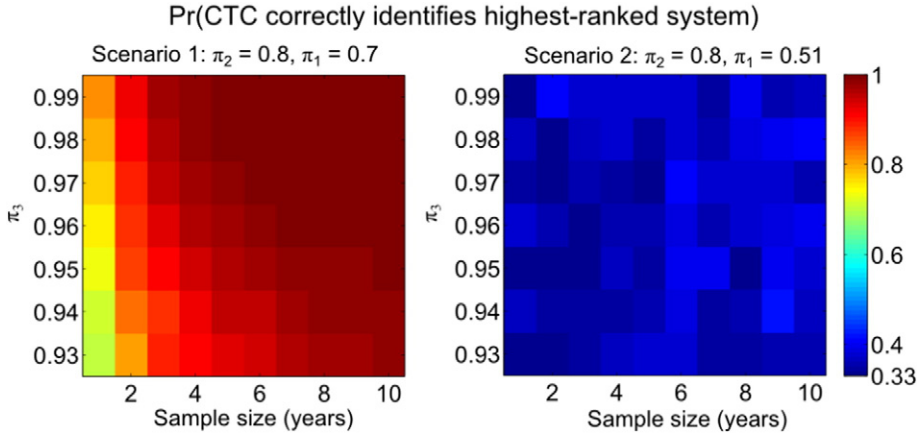

Fig. 3. Probability of CTC correctly identifying the highest-ranked measurement system, for varying sample sizes and measurement system balanced accuracies. In scenario 1 (left), all three measurement systems display moderate to high balanced accuracies. As the balanced accuracy of the highest-ranked system (system 3) is increased and/or the sample size is increased, the probability of CTC correctly identifying the highest-ranked measurement system increases. In scenario 2 (right), measurement system 1 has a very low balanced accuracy. As a result, CTC struggles to identify the highest-ranked measurement system, even for large sample sizes.

(Rignot et al., 1994; Rignot & Way, 1994; Way, Zimmermann, Rignot, McDonald, & Oren, 1997). These continuous measurements can also be thresholded into FT states and quantitatively compared in binary space, usually in terms of classification accuracy μ_i (Colliander et al., 2012; Kim, Kimball, McDonald, & Glassy, 2011; Podest, McDonald, & Kimball, 2014; Zwieback, Bartsch, Melzer, & Wagner, 2012) or sensitivity ψ_i (Zhang, Armstrong, & Smith, 2003). This approach assumes that only the satellite product contains errors. However, air- and even soil-temperatures are only indirectly related to landscape FT state, particularly during the transition seasons, so FT products generated from these proxies will inevitably contain their own errors. In addition, spatial representativeness errors induced from point versus area comparisons also adversely influence confidence in deriving FT 'truth' from in-situ measurements.

4.1. Data

As a proof of concept, we apply CTC to assess the relative performance of three FT products at 33 study sites located across Saskatchewan and Alberta, Canada over the period August 28, 2011–May 25, 2014. The "old black spruce" Boreal Ecosystem Research and Monitoring Site (BERMS) is located in the southern boreal forest in central Saskatchewan. It is flat and homogeneously covered by black spruce forest with soil organic layers 20–30 cm deep (Gower et al., 1997). The 32 other stations are part of a large network maintained by Alberta Agriculture and Rural Development, AgroClimatic Information Service (ACIS, http://agriculture.alberta.ca/acis/) and are situated in open prairie locations.

The first type of FT estimates are derived from in-situ station air temperature measurements (T_a) at each of the 33 stations according to

$$X_1 = \begin{cases} 1, & \text{if } T_a \le 0^\circ\text{C} \\ -1, & \text{if } T_a > 0^\circ\text{C} \end{cases} \tag{15}$$

Air temperature is routinely measured at weather and micrometeorological stations. While soil temperature measurements could also be used to determine the FT state, air temperature has been widely used for FT validation (e.g., Kim, Kimball, Zhang, & McDonald, 2012) and serves as a proxy for wet snow in spring (which is important for the satellite derived FT estimates described next). The use of air temperature also mitigates uncertainty in the spatial representativeness of point soil moisture and temperature measurements due to the high degree of local scale variability (Famiglietti, Ryu, Berg, Rodell, & Jackson, 2008). Furthermore, soil temperature is not available at most weather stations. We only consider early morning (~6 am) measurements where the soil and overlying air are likely in an isothermal state.

The second type of FT estimates are obtained from passive microwave satellite observations. Launched in June 2011, the NASA/SAC-D Aquarius satellite provides collocated active (scatterometer) and passive (radiometer) L-band observations of the Earth's surface at three incidence angles: 29.2°, 38.4° and 46.3° (Le Vine, Lagerloef, Colomb, Yueh, & Pellerano, 2007). It has a repeat time of seven days, crossing the equator twice daily, and an average footprint size of approximately 100 km. While its principal aim is to retrieve sea surface salinity, Aquarius observations have demonstrated utility for estimating large-scale land-surface properties (McColl, Entekhabi and Piles, 2014), including FT state (Brucker, Dinnat, & Koenig, 2014; Roy et al., in press). In this study, we use a weekly, gridded Aquarius radiometer brightness temperature product (Brucker et al., 2014), using beam 2 (38.4° incidence angle; 84 km × 120 km resolution), which is closest to the SMAP incidence angle of 40°. To obtain a FT estimate, we first define a normalized polarization ratio (Choudhury, 1989)

$$NPR = \frac{T_{BV} - T_{BH}}{T_{BV} + T_{BH}} \tag{16}$$

where T_{BV} and T_{BH} are the vertically- and horizontally-polarized brightness temperatures (K) measured by the Aquarius radiometer for the descending (~6 am) overpass. We then define the relative frost factor (Roy et al., in press)

$$FF_{\text{rel}}(t) = \frac{NPR(t) - NPR_{\text{frozen}}}{NPR_{\text{thawed}} - NPR_{\text{frozen}}} \tag{17}$$

where NPR_{frozen} is approximated by the mean of the five lowest winter (January and February) NPR values over the study period; and NPR_{thawed} by the mean of the five highest summer (July and August) NPR values over the study period. The FT prediction X_1 is then given by (Roy et al., in press)

$$X_2 = \begin{cases} 1, & \text{if } FF_{\text{rel}}(t) \le \Delta \\ -1, & \text{if } FF_{\text{rel}}(t) > \Delta \end{cases} \tag{18}$$

where $X_1 = 1$ means 'frozen' and $X_1 = -1$ means 'thawed'. The threshold parameter Δ can be fixed at 0.5 or optimized as described in Kim et al. (2012); in this study, we use the optimized values of 0.46 for BERMS and 0.39 for the prairie sites, determined in a previous study (Roy et al., in press). The seasonal threshold approach is similar to the SMAP baseline FT algorithm (Dunbar et al., 2015). Satellite observations must be considered as measurements of only the upper soil layer FT state, and will also contain contributions from vegetation and interactions with snow (Roy et al., in press).

Finally, a third FT estimate is obtained from soil temperatures (T_s) from the Canadian Meteorological Centre (CMC) surface analysis (Bélair, Brown, Mailhot, Bilodeau, & Crevier, 2003; Bélair, Crevier, Mailhot, Bilodeau, & Delage, 2003) according to

$$X_3 = \begin{cases} 1, & \text{if } T_s \le 0^\circ\text{C} \\ -1, & \text{if } T_s > 0^\circ\text{C} \end{cases} \tag{19}$$

While this product is an analysis that includes station air-temperatures (in addition to a land surface model), it is independent of the stations described previously, which are from research sites. The model's spatial resolution is 0.22° (Bélair, Crevier, et al., 2003).

Example time series of FF_{rel}, T_a and T_s and their corresponding FT products are shown in Fig. 4 for the BERMS site in Saskatchewan. Each product follows a strong seasonal cycle, and suggests that the proposed data model (Fig. 1d) is a reasonable choice for these data. In particular, the variance of each product appears to approach zero during winter/summer, and reach a maximum during the spring/autumn transition periods. Similar seasonal cycles are observed at the other sites used in this study (not shown).

The three datasets were temporally- and spatially-collocated at each site, using nearest-neighbor sampling for the spatial collocation, and the in-situ air temperatures as the temporal reference with a maximum temporal collocation window of 1 day. This procedure resulted in between 123 and 144 sample triplets, depending on the site (because of gaps in the Aquarius or in-situ data). CTC was applied to the sample triplets at each site to obtain performance rankings. We performed bootstrapping (Efron & Tibshirani, 1994) using M (= 1000) replicates, to quantify the impact of sampling error on the estimated rankings. Briefly, for a given site with N sample triplets, N sample triplets are randomly drawn from the N available samples *with replacement*. CTC is then performed on this bootstrapped sample, and the process is repeated M times, producing M performance rankings for the site. If sampling error is small, the CTC performance rankings will be mostly consistent across the M replicates; if the sample size is too small and sampling error is larger, the M performance rankings will be more variable.

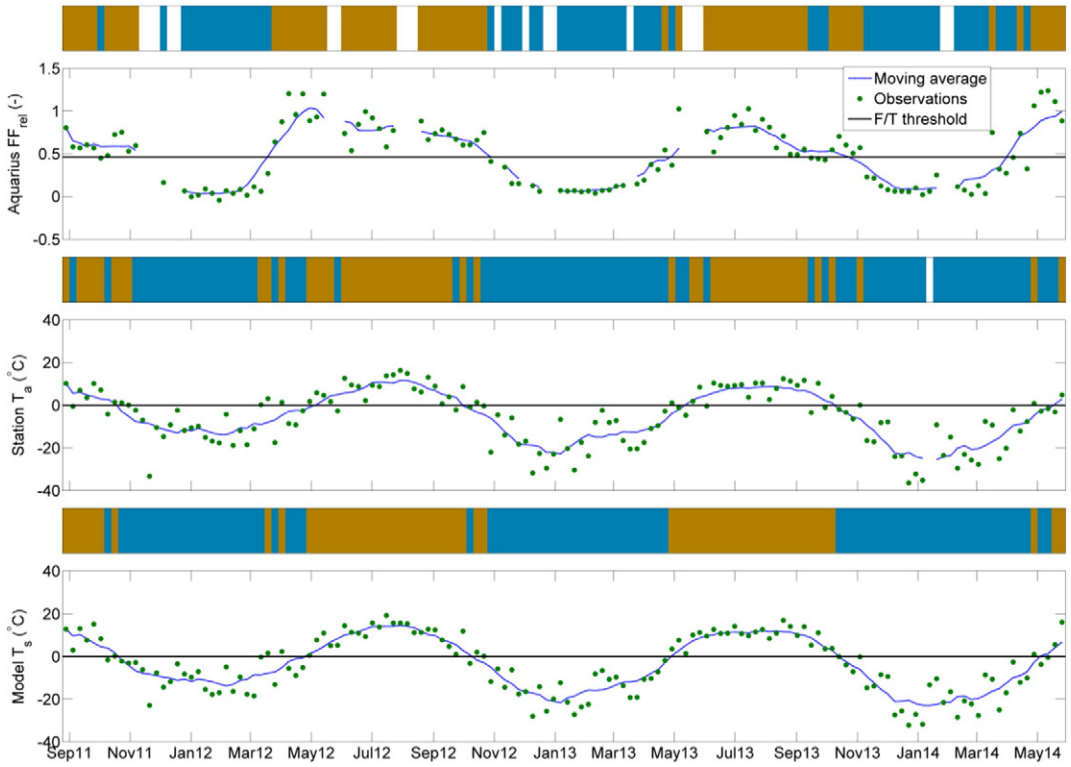

Fig. 4. Time series (points) at BERMS of relative frost factors obtained from Aquarius (top), in-situ station air temperatures T_a (middle) and CMC surface analysis soil temperatures T_s (bottom), with corresponding 4-week moving averages and FT thresholds. Boxes above the time series plots show the inferred FT state for each measurement system, where light brown, blue and white boxes signify thawed, frozen and missing data, respectively.

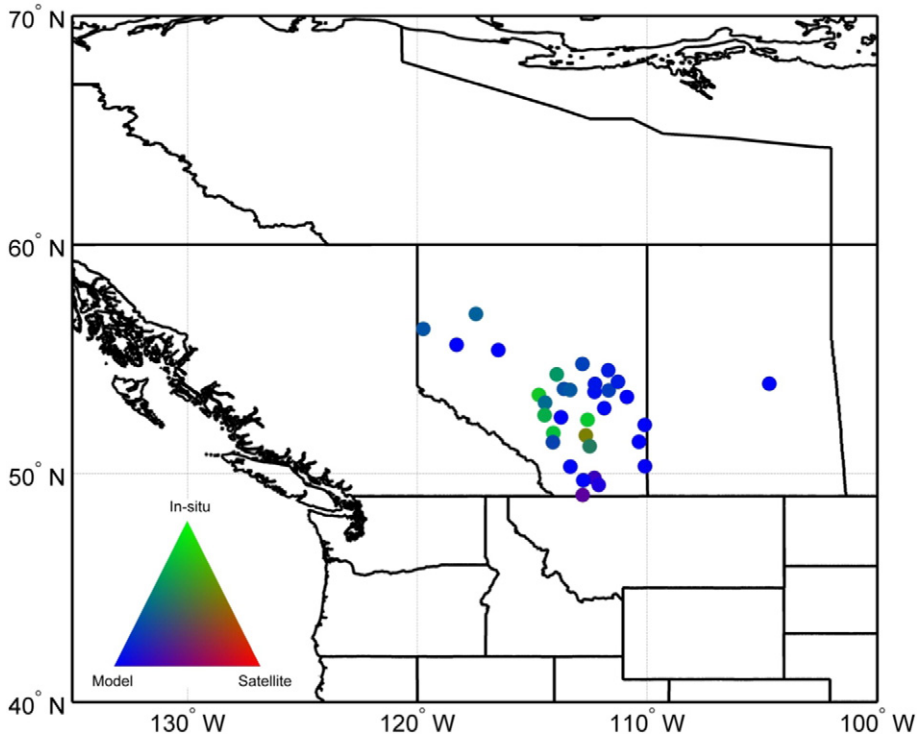

Fig. 5. Map of the measurement system ranked first at each site using CTC. Since the top-ranked measurement system may vary across the M (= 1000) bootstrap replicates at any given site, we calculate the proportion of bootstrap replicates ranking the satellite, in-situ and surface analysis measurements first, and map these to a red–green–blue color space, respectively. For example, if the satellite is ranked the highest in all 1000 bootstrapped performance rankings at a site, it is colored red; if the satellite is ranked first in 50% of the bootstrapped rankings, and the surface analysis is ranked first in the other 50%, it is colored purple.

5. Results and discussion

Fig. 5 maps the measurement systems with the highest CTC performance ranking at each site. The CMC surface analysis is confidently ranked the highest at most sites, including the forested site at BERMS (the easternmost site in Fig. 5). However, at several sites in southeastern Alberta, the in-situ measurements are ranked the highest. The satellite retrievals are confidently ranked second at almost all sites, although the in-situ measurements are ranked second at the forested BERMS site (Fig. 6). Finally, the in-situ air temperature measurements are ranked third at most sites (Fig. 7), although the surface analysis is ranked third at several sites in southeastern Alberta; and at BERMS, the satellite is ranked third (with considerable uncertainty).

Overall, while there is some variability across sites, at most sites, the surface analysis is ranked first, the satellite second, and the in-situ air temperature product third. These results can be explained by a combination of factors. First, the ranking is likely partly a function of the different scales of the FT measurements (~1° for the satellite product, 0.22° for the model product, and point-scale for the in-situ product). Therefore, so-called representativeness errors will be significant, particularly for the in-situ product (Miralles et al., 2010) and, to a lesser extent, the satellite product (Gruber et al., in press). Indeed, if there is significant heterogeneity in the FT field, the top ranking of the surface analysis could be explained entirely by its intermediate scale between those of the satellite and in-situ products: as noted earlier, CTC rewards product i for covarying with products j and k, and even more so if products j and k covary little.

Second, the ranking is likely also due to differences in measurement accuracy of the different products. The fact that the satellite measurements are generally not ranked first is consistent with the multiple sources of uncertainty with respect to surface influences on permittivity (snow, vegetation and soil) and hence the brightness temperature time series: for example, the dense forest at BERMS, crop growth over the prairie sites, topography, and/or the presence of subgrid wetlands or waterbodies (Du et al., 2014; Podest et al., 2014). And while representativeness errors are undoubtedly an important factor in the low ranking of the in-situ station product, even point measurements of air temperature contain information about air temperatures across a much wider region (this would not be true for soil temperatures, which are highly variable at small scales). Hence, perhaps the use of station air temperatures rather than soil temperatures introduces significant errors into the in-situ product, beyond representativeness errors. For instance, during spring, there can be a pronounced time lag between snowmelt onset and soil temperature response (i.e., air temperatures rise above freezing, the snow begins to melt but the soil remains frozen). Roy et al. (in press) show a spread of up to four weeks in the estimate of spring thaw onset from soil temperature, air temperature and satellite-derived land surface temperature datasets.

Our analysis is subject to several limitations. While the sample sizes used are comparable with many other published TC studies, sampling error is large enough to obscure complete performance rankings at some sites (for instance, it is unclear which system is ranked third at BERMS). This problem is not unique to CTC, and would apply to any attempt to validate the available satellite FT estimates, for instance, by calculating the accuracy relative to station measurements. The available sample size will increase as the L-band satellite record grows, and as the frequency of observations increases in future satellite missions (from Aquarius's 7-day repeat time to SMAP's 3-day repeat time). Sampling issues could be mitigated in future validation studies by conducting CTC on spatial samples, rather than temporal samples; such an analysis would require the in-situ station measurements to be replaced with another product with greater spatial coverage, for instance, a higher-frequency radar product. Some of the standard problems of TC may also apply. A previous study has shown that the assumptions of classical TC are often violated when applied to soil moisture (Yilmaz & Crow, 2014). It is possible that (R1*) is violated in our analysis, for example, due to representativeness errors (Stoffelen, 1998; Vogelzang et al., 2011), although we note that (R1*) is a substantially weaker assumption than (R1) and (R2), and thus less likely to be violated in general. Furthermore, CTC is inherently more robust to violations of assumptions

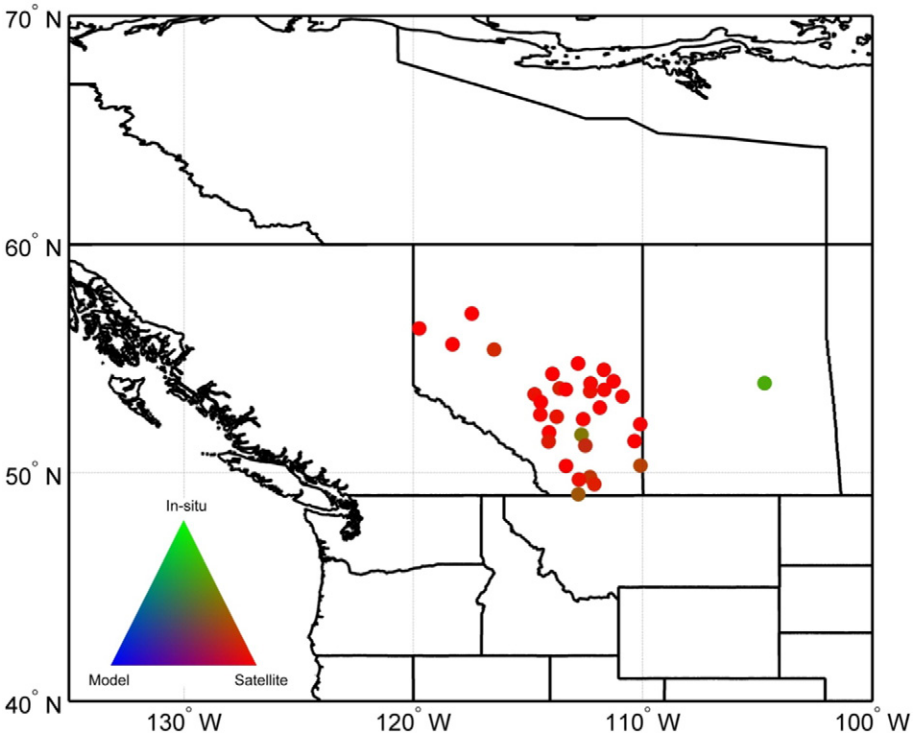

Fig. 6. Similar to Fig. 5 for the measurement system ranked second at each site using CTC.

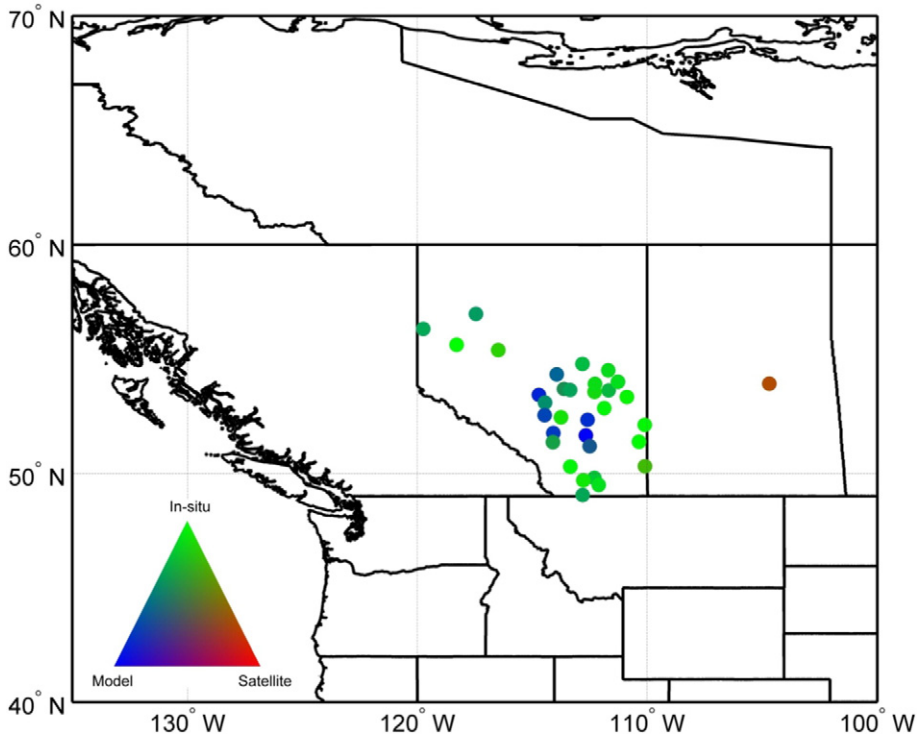

Fig. 7. Similar to Fig. 5 for the measurement system ranked third at each site using CTC.

compared to classical TC since it estimates performance rankings, rather than absolute values of performance metrics: estimates of $\boldsymbol{w}$ (Eq. (14)) may be biased but still lead to the correct ranking.

6. Conclusions

Classical TC is limited to continuous variables and measurement system triplets that have uncorrelated errors and errors independent of the state variable. In this study, we show that, when TC is applied to categorical variables, both assumptions are automatically violated. We introduce CTC, a variant of TC that is compatible with categorical variables. Application of CTC allows relative ranking of three measurement systems in proportion to their balanced accuracies. As an example application, CTC is applied to estimate performance rankings of FT estimates generated from point in-situ air temperatures, a normalized polarization ratio from passive L-band satellite observations, and soil temperatures from an atmospheric model analysis.

CTC analyses can be viewed as complementary to traditional validation studies that treat a measurement system as error-free (e.g., in-situ data). While CTC can only provide performance rankings, not error magnitudes, it can be used to identify the most accurate measurement system at a particular location, which can then be used as the error-free "truth" system in a traditional error characterization study to obtain estimates of error magnitudes. Our results indicate that, for at least some regions, it may be more appropriate to characterize satellite FT errors by comparison with a combination of model-based and in-situ FT estimates, rather than those based solely on in-situ air temperatures.

Some of the problems identified here with applying classical TC to categorical variables also apply to bounded continuous variables, such as soil moisture. While the impacts will be minimal for soil moisture values away from the boundaries, classical TC should be applied with caution in very dry (wet) regions where soil moisture is frequently near its lower (upper) bound, where we expect strong violations of (R1) and (R2).

Code is available for implementing CTC at https://github.com/kaighin.

Acknowledgments

The authors thank Shunli Zhang and Stéphane Bélair (Environment Canada) and Youngwook Kim and John Kimball (University of Montana) for supplying data; and Alexander Gruber and three anonymous reviewers for their constructive feedback. K.A.M. is supported by the NSF Graduate Research Fellowship Program. A.G.K. was supported by a NASA Earth and Space Science Fellowship.

Appendix A. Inapplicability of classical TC for categorical variables

We show that naïve application of standard TC to binary target variables will automatically result in the violation of both key error assumptions. We begin with a standard error model used in triple collocation:

$$X_i = T + \varepsilon_i \tag{A1}$$

Without loss of generality, we choose $X_i, T \in \{-1, 1\}$ (i.e., the result of this derivation is independent of the labels used for the binary variables). We do not use the affine error model, since a linear relationship between observations X_i and target variable t is inappropriate in the binary case. This error model requires that $\mathrm{E}(\varepsilon_i) = 0$, so the error distribution must have the form

$$p_{E_i}(\varepsilon_i) = \begin{cases} \frac{1}{2}P_i, & \text{for } \varepsilon_i = 2 \\ 1-P_i, & \text{for } \varepsilon_i = 0 \\ \frac{1}{2}P_i, & \text{for } \varepsilon_i = -2 \end{cases}, \tag{A2}$$

where p_{E_i} is the error probability mass function (PMF) and P_i is a parameter such that the probability of an error occurring for measurement system i is P_i.

First, we show that assumption (R2) is violated, i.e., $\mathrm{Cov}(\varepsilon_i, T) \neq 0$. When $T=1$, ε_i must equal either 0 or -2 to produce $X_i \in \{-1, 1\}$; similarly, when $T=-1$, ε_i must equal either 0 or 2. Therefore, the PMF of $T\varepsilon_i$ can be written as

$$p_{TE_i}(T\varepsilon_i) = \begin{cases} P_i, & \text{for } T\varepsilon_i = -2 \\ 1-P_i, & \text{for } T\varepsilon_i = 0 \end{cases}, \tag{A3}$$

Therefore, $\mathrm{E}(T\varepsilon_i) = -2P_i$, so $\mathrm{Cov}(\varepsilon_i, T) = \mathrm{E}(T\varepsilon_i) - \mathrm{E}(T)\mathrm{E}(\varepsilon_i) = -2P_i$, since $\mathrm{E}(\varepsilon_i) = 0$. This is non-zero for all non-trivial cases in which $P_i > 0$.

Second, we show that assumption (R1) is violated, i.e., $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) \neq 0$ for $i \neq j$. By the law of total covariance, we have

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \mathrm{E}\big(\mathrm{Cov}(\varepsilon_i, \varepsilon_j | T)\big) + \mathrm{Cov}\big(\mathrm{E}(\varepsilon_i | T), \mathrm{E}(\varepsilon_j | T)\big). \tag{A4}$$

Furthermore, by definition,

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j | T) = \mathrm{E}(\varepsilon_i \varepsilon_j | T) - \mathrm{E}(\varepsilon_i | T)\mathrm{E}(\varepsilon_j | T). \tag{A5}$$

The distribution of the product of the errors $\varepsilon_i \varepsilon_j$ conditioned on T can be written

$$p_{E_iE_j|T}(\varepsilon_i\varepsilon_j | T=1) = p_{E_iE_j|T}(\varepsilon_i\varepsilon_j | T=-1) = \begin{cases} P_iP_j, & \text{for } \varepsilon_i\varepsilon_j = 4 \\ 1-P_iP_j, & \text{for } \varepsilon_i\varepsilon_j = 0 \end{cases} \tag{A6}$$

Therefore, $\mathrm{E}(\varepsilon_i\varepsilon_j | T=1) = \mathrm{E}(\varepsilon_i\varepsilon_j | T=-1) = 4P_iP_j$. The distribution of the errors ε_i conditioned on T can be written as

$$p_{E_i|T}(\varepsilon_i | T=1) = \begin{cases} P_i, & \text{for } \varepsilon_i = -2 \\ 1-P_i, & \text{for } \varepsilon_i = 0 \end{cases} \tag{A7}$$

and

$$p_{E_i|T}(\varepsilon_i | T=-1) = \begin{cases} P_i, & \text{for } \varepsilon_i = 2 \\ 1-P_i, & \text{for } \varepsilon_i = 0 \end{cases} \tag{A8}$$

Therefore, $\mathrm{E}(\varepsilon_i | T) = -2P_iT$ and $\mathrm{E}(\varepsilon_i | T)\mathrm{E}(\varepsilon_j | T) = 4P_iP_j$. Substituting these results into Eq. (A5) and taking the expectation, we obtain

$$\mathrm{E}\big(\mathrm{Cov}(\varepsilon_i, \varepsilon_j | T)\big) = \mathrm{E}\big(4P_iP_j - 4P_iP_j\big) = 0, \tag{A9}$$

Also,

$$\begin{aligned} \mathrm{Cov}\big(\mathrm{E}(\varepsilon_i | T), \mathrm{E}(\varepsilon_j | T)\big) &= \mathrm{Cov}\big(-2P_iT, -2P_jT\big) \\ &= \mathrm{E}\Big(4P_iP_jT^2\Big) - 4P_iP_j\mathrm{E}(T)^2 = 4P_iP_j\mathrm{Var}(T). \end{aligned} \tag{A10}$$

Therefore, substituting Eqs. (A9) and (A10) into Eq. (A4) yields the result $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 4P_iP_j\mathrm{Var}(T) > 0$ for all non-trivial cases where $P_i > 0, P_j > 0$ and $\mathrm{Var}(T) > 0$.

Appendix B. Non-zero error means

It is shown in Appendix A that the assumption $\mathrm{E}(\varepsilon_i) = 0$ ensures that the error distribution is symmetric and that $\mathrm{E}(\varepsilon_i | T) = -2P_iT$. Then, by the law of total expectation, we have

$$0 = \mathrm{E}(\varepsilon_i) = \mathrm{E}\big(\mathrm{E}(\varepsilon_i | T)\big) = \mathrm{E}(-2P_iT) = -2P_iE(T), \tag{B1}$$

so $E(T) = 0$. Furthermore, it implies that $\mathrm{E}(X_i) = \mathrm{E}(T) + \mathrm{E}(\varepsilon_i) = 0$.

Now consider the more general case where ε_i need not have zero mean. Then

$$p_{E_i|T}(\varepsilon_i | T=1) = \begin{cases} 1-\psi_i, & \text{for } \varepsilon_i = -2 \\ \psi_i, & \text{for } \varepsilon_i = 0 \end{cases}. \tag{B2}$$

and

$$p_{E_i|T}(\varepsilon_i | T=-1) = \begin{cases} 1-\eta_i, & \text{for } \varepsilon_i = 2 \\ \eta_i, & \text{for } \varepsilon_i = 0 \end{cases} \tag{B3}$$

Therefore $\mathrm{E}(\varepsilon_i | T=1) = -2(1-\psi_i)$, $\mathrm{E}(\varepsilon_i | T=-1) = 2(1-\eta_i)$ and so

$$\begin{aligned} \mathrm{E}(\varepsilon_i) &= \mathrm{E}\big(\mathrm{E}(\varepsilon_i | T)\big) = 2\big(1-\eta_i\big)p_T(T=-1) - 2(1-\psi_i)p_T(T=1) \\ &= 2\big((\psi_i + \eta_i - 2)p_T(T=1) + 1 - \eta_i\big). \end{aligned} \tag{B4}$$

Appendix C. Generalization of P14 to time series

This result generalizes the relation between covariances and balanced accuracies given in P14 to time series.

P14 show, for the stationary case, that

$$\mathrm{Cov}(X_i, X_j) = \begin{cases} 1 - E(X_i)^2, & \text{for } i = j \\ Var(T)(2\pi_i - 1)(2\pi_j - 1), & \text{otherwise} \end{cases} \tag{C1}$$

We refer the reader to P14 for details of the derivation. For the non-stationary case, $\mathrm{Var}(T)$ varies in time and hence across samples, so this result only holds when conditioned on time t, i.e.,

$$\mathrm{Cov}(X_i, X_j | t) = \begin{cases} 1 - E(X_i|t)^2, & \text{for } i = j \\ 4p(t)(1-p(t))(2\pi_i - 1)(2\pi_j - 1), & \text{otherwise} \end{cases} \tag{C2}$$

since $\mathrm{Var}(T|t) = 4p(t)(1-p(t))$ where $p(t) \equiv p_T(T=1|t)$. We seek the unconditional covariance $\mathrm{Cov}(X_i, X_j)$ for the non-stationary case. By the law of total covariance,

$$\mathrm{Cov}(X_i, X_j) = \mathrm{E}\big(\mathrm{Cov}(X_i, X_j)|t\big) + \mathrm{Cov}\big(\mathrm{E}(X_i|t), \mathrm{E}(X_j|t)\big) \tag{C3}$$

Starting with the $i=j$ case in Eq. (C2), we have

$$\mathrm{E}\big(\mathrm{Cov}(X_i, X_j)|t\big) = 1 - \mathrm{E}\Big(\mathrm{E}(X_i|t)^2\Big) \tag{C4}$$

and

$$\mathrm{Cov}\big(E(X_i|t), E(X_j|t)\big) = \mathrm{E}\Big(E(X_i|t)^2\Big) - \mathrm{E}(\mathrm{E}(X_i|t))^2 \tag{C5}$$

by definition. Substituting Eqs. (C4) and (C5) into Eq. (C3) gives

$$\mathrm{Cov}(X_i, X_j) = 1 - \mathrm{E}(\mathrm{E}(X_i|t))^2 \tag{C6}$$

Now we consider the case where $i \neq j$ in Eq. (C2). First, note that

$$\begin{aligned} \mathrm{E}(X_i|t) &= \mathrm{E}(T|t) + \mathrm{E}(\varepsilon_i|t) \\ &= 2(\psi_i + \eta_i - 1)p(t) + 1 - 2\eta_i \end{aligned} \tag{C7}$$

obtained by combining the results that $\mathrm{E}(T|t) = 2p(t) - 1$ (by definition) and Eq. (B4).

Second, note that

$$\begin{aligned} \mathrm{E}\big(\mathrm{E}(X_i|t)\mathrm{E}(X_j|t)\big) = {} & 4(\psi_i + \eta_i - 1)\Big(\psi_j + \eta_j - 1\Big)\mathrm{E}\Big(p(t)^2\Big) \\ & + 2\Big((\psi_i + \eta_i - 1)\Big(1 - 2\eta_j\Big) + \Big(\psi_j + \eta_j - 1\Big)(1 - 2\eta_i)\Big)\mathrm{E}(p(t)) \\ & + (1 - 2\eta_i)\Big(1 - 2\eta_j\Big) \end{aligned} \tag{C8}$$

using Eq. (C7).

Similarly, we have

$$\begin{aligned}\mathrm{E}(\mathrm{E}(X_i|t))\mathrm{E}(\mathrm{E}(X_j|t)) &= 4(\psi_i+\eta_i-1)\left(\psi_j+\eta_j-1\right)\mathrm{E}(p(t))^2 \\ &+ 2\left((\psi_i+\eta_i-1)\left(1-2\eta_j\right)+\left(\psi_j+\eta_j-1\right)(1-2\eta_i)\right)\mathrm{E}(p(t)) \\ &+ (1-2\eta_i)\left(1-2\eta_j\right)\end{aligned} \tag{C9}$$

Therefore, using Eqs. (C8) and (C9), we have

$$\begin{aligned}\mathrm{Cov}\left(\mathrm{E}(X_i|t),\mathrm{E}(X_j|t)\right) &= \mathrm{E}\left(\mathrm{E}(X_i|t)\mathrm{E}(X_j|t)\right)-\mathrm{E}(\mathrm{E}(X_i|t))\mathrm{E}\left(\mathrm{E}(X_j|t)\right) \\ &= 4\left(\mathrm{E}\left(p(t)^2\right)-\mathrm{E}(p(t))^2\right)(\psi_i+\eta_i-1)\left(\psi_j+\eta_j-1\right) \\ &= 4\left(\mathrm{E}\left(p(t)^2\right)-\mathrm{E}(p(t))^2\right)(2\pi_i-1)(2\pi_j-1)\end{aligned} \tag{C10}$$

Therefore, by substituting Eqs. (C2) and (C10) into Eq. (C3) and rearranging, we obtain the unconditional covariance for the case where $i \neq j$ and have

$$\mathrm{Cov}(X_i,X_j) = \begin{cases} 1-\mathrm{E}(\mathrm{E}(X_i|t))^2, & \text{for } i=j \\ 4\mathrm{E}(p(t))(1-\mathrm{E}(p(t)))(2\pi_i-1)(2\pi_j-1), & \text{otherwise}\end{cases} \tag{C11}$$

Note that this reduces to Eq. (C1), the stationary case given in P14, when T is not dependent on time t.

References

Ackerman, S.A., Strabala, K.I., Menzel, W.P., Frey, R.A., Moeller, C.C., & Gumley, L.E. (1998). Discriminating clear sky from clouds with MODIS. *Journal of Geophysical Research – Atmospheres*, *103*, 32141–32157. http://dx.doi.org/10.1029/1998JD200032.

Alemohammad, S.H., McColl, K.A., Konings, A.G., Entekhabi, D., & Stoffelen, A. (2015). Characterization of precipitation product errors across the United States using multiplicative triple collocation. *Hydrology and Earth System Sciences*, *19*, 3489–3503. http://dx.doi.org/10.5194/hess-19-3489-2015.

Bélair, S., Brown, R., Mailhot, J., Bilodeau, B., & Crevier, L. -P. (2003a). Operational implementation of the ISBA land surface scheme in the Canadian regional weather forecast model. Part II: Cold season results. *Journal of Hydrometeorology*, *4*, 371–386.

Bélair, S., Crevier, L. -P., Mailhot, J., Bilodeau, B., & Delage, Y. (2003b). Operational implementation of the ISBA land surface scheme in the Canadian regional weather forecast model. Part I: Warm season results. *Journal of Hydrometeorology*, *4*, 352–370.

Betts, A.K., Viterbo, P., Beljaars, A.C.M., & van den Hurk, B.J.J.M. (2001). Impact of BOREAS on the ECMWF forecast model. *Journal of Geophysical Research – Atmospheres*, *106*, 33593–33604. http://dx.doi.org/10.1029/2001JD900056.

Brucker, L., Dinnat, E.P., & Koenig, L.S. (2014). Weekly gridded Aquarius L-band radiometer/scatterometer observations and salinity retrievals over the polar regions — Part 1: Product description. *The Cryosphere*, *8*, 905–913. http://dx.doi.org/10.5194/tc-8-905-2014.

Choudhury, B.J. (1989). Monitoring global land surface using Nimbus-7 37 GHz data theory and examples. *International Journal of Remote Sensing*, *10*, 1579–1605. http://dx.doi.org/10.1080/01431168908903993.

Colliander, A., McDonald, K., Zimmermann, R., Schroeder, R., Kimball, J.S., & Njoku, E.G. (2012). Application of QuikSCAT backscatter to SMAP validation planning: Freeze/thaw state over ALECTRA sites in Alaska from 2000 to 2007. *IEEE Transactions on Geoscience and Remote Sensing*, *50*, 461–468. http://dx.doi.org/10.1109/TGRS.2011.2174368.

D'Odorico, P., Gonsamo, A., Pinty, B., Gobron, N., Coops, N., Mendez, E., & Schaepman, M.E. (2014). Intercomparison of fraction of absorbed photosynthetically active radiation products derived from satellite data over Europe. *Remote Sensing of Environment*, *142*, 141–154. http://dx.doi.org/10.1016/j.rse.2013.12.005.

Draper, C., Reichle, R., de Jeu, R., Naeimi, V., Parinussa, R., & Wagner, W. (2013). Estimating root mean square errors in remotely sensed soil moisture over continental scale domains. *Remote Sensing of Environment*, *137*, 288–298. http://dx.doi.org/10.1016/j.rse.2013.06.013.

Du, J., Kimball, J.S., Azarderakhsh, M., Dunbar, R.S., Moghaddam, M., & McDonald, K.C. (2014). Classification of Alaska spring thaw characteristics using satellite L-band radar remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*. http://dx.doi.org/10.1109/TGRS.2014.2325409 (Early Access Online).

Dunbar, S., Xu, X., Colliander, A., Derksen, C., Kimball, J., McDonald, K., ... Podest, E. (2015). *Soil moisture active/passive L3 freeze–thaw algorithm theoretical basis document.*

Efron, B., & Tibshirani, R.J. (1994). *An introduction to the bootstrap.* CRC Press.

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

Entekhabi, D., Reichle, R.H., Koster, R.D., & Crow, W.T. (2010b). Performance metrics for soil moisture retrievals and application requirements. *Journal of Hydrometeorology*, *11*, 832–840. http://dx.doi.org/10.1175/2010JHM1223.1.

Famiglietti, J.S., Ryu, D., Berg, A., Rodell, M., & Jackson, T. (2008). Field observations of soil moisture variability across scales. *Water Resources Research*, *44*.

Fang, H., Wei, S., Jiang, C., & Scipal, K. (2012). Theoretical uncertainty analysis of global MODIS, CYCLOPES, and GLOBCARBON LAI products using a triple collocation method. *Remote Sensing of Environment*, *124*, 610–621. http://dx.doi.org/10.1016/j.rse.2012.06.013.

Farhadi, L., Reichle, R.H., De Lannoy, G.J.M., & Kimball, J.S. (2014). Assimilation of freeze/thaw observations into the NASA catchment land surface model. *Journal of Hydrometeorology*. http://dx.doi.org/10.1175/JHM-D-14-0065.1.

Friedl, M.A., McIver, D.K., Hodges, J.C.F., Zhang, X.Y., Muchoney, D., Strahler, A.H., ... Schaaf, C. (2002). Global land cover mapping from MODIS: Algorithms and early results. *Remote Sensing of Environment*, *83*, 287–302. http://dx.doi.org/10.1016/S0034-4257(02)00078-0 (The Moderate Resolution Imaging Spectroradiometer (MODIS): A new generation of Land Surface Monitoring).

Goulden, M.L., Wofsy, S.C., Harden, J.W., Trumbore, S.E., Crill, P.M., Gower, S.T., ... Munger, J.W. (1998). Sensitivity of boreal forest carbon balance to soil thaw. *Science*, *279*, 214–217. http://dx.doi.org/10.1126/science.279.5348.214.

Gower, S.T., Vogel, J.G., Norman, J.M., Kucharik, C.J., Steele, S.J., & Stow, T.K. (1997). Carbon distribution and aboveground net primary production in aspen, jack pine, and black spruce stands in Saskatchewan and Manitoba, Canada. *Journal of Geophysical Research – Atmospheres*, *102*, 29029–29041. http://dx.doi.org/10.1029/97JD02317.

Gruber, A., Su, C. -H., Zwieback, S., Crow, W., Dorigo, W., & Wagner, W. (2016). Recent advances in (soil moisture) triple collocation analysis. *International Journal of Applied Earth Observation and Geoinformation*. http://dx.doi.org/10.1016/j.jag.2015.09.002 (in press).

Janssen, P.A.E.M., Abdalla, S., Hersbach, H., & Bidlot, J. -R. (2007). Error estimation of buoy, satellite, and model wave height data. *Journal of Atmospheric and Oceanic Technology*, *24*, 1665–1677. http://dx.doi.org/10.1175/JTECH2069.1.

Kapteyn, A., & Ypma, J.Y. (2007). Measurement error and misclassification: A comparison of survey and administrative data. *Journal of Labor Economics*, *25*, 513–551. http://dx.doi.org/10.1086/513298.

Kim, Y., Kimball, J., Zhang, K., & McDonald, K. (2012). Satellite detection of increasing Northern Hemisphere non-frozen seasons from 1979 to 2008: Implications for regional vegetation growth. *Remote Sensing of Environment*, *121*, 472–487.

Kim, Y., Kimball, J.S., McDonald, K.C., & Glassy, J. (2011). Developing a global data record of daily landscape freeze/thaw status using satellite passive microwave remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, *49*, 949–960. http://dx.doi.org/10.1109/TGRS.2010.2070515.

Le Vine, D.M., Lagerloef, G.S.E., Colomb, F.R., Yueh, S.H., & Pellerano, F.A. (2007). Aquarius: An instrument to monitor sea surface salinity from space. *IEEE Transactions on Geoscience and Remote Sensing*, *45*, 2040–2050. http://dx.doi.org/10.1109/TGRS.2007.898092.

McColl, K.A., Entekhabi, D., & Piles, M. (2014a). Uncertainty analysis of soil moisture and vegetation indices using aquarius scatterometer observations. *IEEE Transactions on Geoscience and Remote Sensing*, *52*, 4259–4272. http://dx.doi.org/10.1109/TGRS.2013.2280701.

McColl, K.A., Vogelzang, J., Konings, A.G., Entekhabi, D., Piles, M., & Stoffelen, A. (2014b). Extended triple collocation: Estimating errors and correlation coefficients with respect to an unknown target. *Geophysical Research Letters*, *41*, 2014GL061322. http://dx.doi.org/10.1002/2014GL061322.

Metternicht, G., Hurni, L., & Gogu, R. (2005). Remote sensing of landslides: An analysis of the potential contribution to geo-spatial systems for hazard assessment in mountainous environments. *Remote Sensing of Environment*, *98*, 284–303. http://dx.doi.org/10.1016/j.rse.2005.08.004.

Miralles, D.G., Crow, W.T., & Cosh, M.H. (2010). Estimating spatial sampling errors in coarse-scale soil moisture estimates derived from point-scale observations. *Journal of Hydrometeorology*, *11*, 1423–1429. http://dx.doi.org/10.1175/2010JHM1285.1.

O'Carroll, A.G., Eyre, J.R., & Saunders, R.W. (2008). Three-way error analysis between AATSR, AMSR-E, and in situ sea surface temperature observations. *Journal of Atmospheric and Oceanic Technology*, *25*, 1197–1207. http://dx.doi.org/10.1175/2007JTECHO542.1.

Parisi, F., Strino, F., Nadler, B., & Kluger, Y. (2014). Ranking and combining multiple predictors without labeled data. *Proceedings of the National Academy of Sciences*, *111*, 1253–1258.

Podest, E., McDonald, K.C., & Kimball, J.S. (2014). Multisensor microwave sensitivity to freeze/thaw dynamics across a complex boreal landscape. *IEEE Transactions on Geoscience and Remote Sensing*, *52*, 6818–6828. http://dx.doi.org/10.1109/TGRS.2014.2303635.

Rautiainen, K., Parkkinen, T., Lemmetyinen, J., Schwank, M., Wiesmann, A., Ikonen, J., ... Pulliainen, J. (2016). SMOS prototype algorithm for detecting autumn soil freezing. *Remote Sensing of Environment*. http://dx.doi.org/10.1016/j.rse.2016.01.012.

Rignot, E., & Way, J.B. (1994). Monitoring freeze—thaw cycles along North-South Alaskan transects using ERS-1 SAR. *Remote Sensing of Environment*, *49*, 131–137. http://dx.doi.org/10.1016/0034-4257(94)90049-3.

Rignot, E., Way, J.B., McDonald, K., Viereck, L., Williams, C., Adams, P., ... Shi, J. (1994). Monitoring of environmental conditions in Taiga forests using ERS-1 SAR. *Remote Sensing of Environment*, *49*, 145–154. http://dx.doi.org/10.1016/0034-4257(94)90051-5.

Roebeling, R.A., Wolters, E.L.A., Meirink, J.F., & Leijnse, H. (2012). Triple collocation of summer precipitation retrievals from SEVIRI over Europe with gridded rain gauge and weather radar data. *Journal of Hydrometeorology*, *13*, 1552–1566. http://dx.doi.org/10.1175/JHM-D-11-089.1.

Roy, A., Royer, A., Derksen, C., Brucker, L., Langlois, A., Mialon, A., & Kerr, Y. (2016). Evaluation of spaceborne L-band radiometer measurements for terrestrial freeze/

thaw retrievals in Canada. *Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (in press).

Roy, D.P., Boschetti, L., Justice, C.O., & Ju, J. (2008). The collection 5 MODIS burned area product — Global evaluation by comparison with the MODIS active fire product. *Remote Sensing of Environment*, *112*, 3690–3707. http://dx.doi.org/10.1016/j.rse.2008.05.013.

Stoffelen, A. (1998). Toward the true near-surface wind speed: Error modeling and calibration using triple collocation. *Journal of Geophysical Research*, *103*, 7755–7766.

Vogelzang, J., Stoffelen, A., Verhoef, A., & Figa-Saldaña, J. (2011). On the quality of high-resolution scatterometer winds. *Journal of Geophysical Research*, *116*. http://dx.doi.org/10.1029/2010JC006640.

Way, J., Paris, J., Kasischke, E., Slaughter, C., Viereck, L., Christensen, N., ... Weber, J. (1990). The effect of changing environmental conditions on microwave signatures of forest ecosystems: Preliminary results of the March 1988 Alaskan aircraft SAR experiment. *International Journal of Remote Sensing*, *11*, 1119–1144. http://dx.doi.org/10.1080/01431169008955084.

Way, J., Zimmermann, R., Rignot, E., McDonald, K., & Oren, R. (1997). Winter and spring thaw as observed with imaging radar at BOREAS. *Journal of Geophysical Research – Atmospheres*, *102*, 29673–29684. http://dx.doi.org/10.1029/96JD03878.

Wegmüller, U. (1990). The effect of freezing and thawing on the microwave signatures of bare soil. *Remote Sensing of Environment*, *33*, 123–135. http://dx.doi.org/10.1016/0034-4257(90)90038-N.

Yilmaz, M.T., & Crow, W.T. (2014). Evaluation of assumptions in soil moisture triple collocation analysis. *Journal of Hydrometeorology*. http://dx.doi.org/10.1175/JHM-D-13-0158.1.

Zhang, T., Armstrong, R.L., & Smith, J. (2003). Investigation of the near-surface soil freeze–thaw cycle in the contiguous United States: Algorithm development and validation. *Journal of Geophysical Research – Atmospheres*, *108*, 8860. http://dx.doi.org/10.1029/2003JD003530.

Zwieback, S., Bartsch, A., Melzer, T., & Wagner, W. (2012). Probabilistic fusion of–and C-band scatterometer data for determining the freeze/thaw state. *IEEE Transactions on Geoscience and Remote Sensing*, *50*, 2583–2594. http://dx.doi.org/10.1109/TGRS.2011.2169076.