International Journal of Applied Earth Observation and Geoinformation 45 (2016) 200–211

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

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

Recent advances in (soil moisture) triple collocation analysis

A. Gruber[a,*], C.-H. Su[b], S. Zwieback[c], W. Crow[d], W. Dorigo[a,e], W. Wagner[a]

[a] Department of Geodesy and Geoinformation, Vienna University of Technology, Vienna, Austria
[b] Department of Infrastructure Engineering, University of Melbourne, Victoria 3010, Australia
[c] Institute of Environmental Engineering, ETH Zurich, Zurich, Switzerland
[d] USDA Hydrology and Remote Sensing Laboratory, Beltsville, MD, USA
[e] Laboratory of Hydrology and Water Management, Ghent University, Belgium

ARTICLE INFO

Article history:
Received 31 March 2015
Received in revised form
11 September 2015
Accepted 11 September 2015
Available online 28 September 2015

Keywords:
Soil moisture
Error characterization
Validation
Triple collocation

ABSTRACT

To date, triple collocation (TC) analysis is one of the most important methods for the global-scale evaluation of remotely sensed soil moisture data sets. In this study we review existing implementations of soil moisture TC analysis as well as investigations of the assumptions underlying the method. Different notations that are used to formulate the TC problem are shown to be mathematically identical. While 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. Moreover, assumptions, which are often understood as a limitation that is unique to TC analysis are shown to be common also to other conventional performance metrics. Noteworthy advances in TC analysis have been made in the way error estimates are being presented by moving from the investigation of absolute error variance estimates to the investigation of signal-to-noise ratio (SNR) metrics. Here we review existing error presentations and propose the combined investigation of the SNR (expressed in logarithmic units), the unscaled error variances, and the soil moisture sensitivities of the data sets as an optimal strategy for the evaluation of remotely-sensed soil moisture data sets.

1. Introduction

Soil moisture is one of the most important drivers of the hydrological cycle. Therefore, global soil moisture records are needed to study hydrology driven phenomena of the earth system such as climate change, vegetation growth, and many others (Legates et al., 2011). The most important sources for global soil moisture records are microwave radar and radiometer instruments (Liu et al., 2011), and land surface models (Reichle et al., 2002). However, both satellite measurements and model predictions are subject to errors and their correct interpretation and application requires an in-depth understanding of their accuracy.

Triple collocation (TC) analysis is a method for estimating the random error variances of three collocated data sets of the same geophysical variable (Stoffelen, 1998). It does not require the availability of a high-quality reference data set and has therefore evolved as one of the most important evaluation methods in earth observation. In this study we will focus exclusively on the evaluation of remotely sensed soil moisture, even though some of the discussions and findings are of general validity to other variables in hydrometeorology and oceanography (Vogelzang et al., 2011; Caires and Sterl, 2003; Roebeling et al., 2012; Fang et al., 2012).

Since its development in 1998 a host of research has been carried out to investigate the limitations of TC analysis, most of which are related to violations in the underlying assumptions that are made on the structural properties of the considered data sets. However, only a few studies have proposed methods to mitigate the impact of such violations. Moreover, the assumptions made in TC analysis are often considered to be unique to the method, yet most of them are also implicitly made in the application of conventional performance metrics, which has not been explicitly pointed out in existing studies. This study will provide a comprehensive discussion of the assumptions that are made for TC analysis and the impact of possible violations, together with a review of already existing investigations and proposed modifications of the TC model. Also, we will demonstrate the similarity between the assumptions that are made for TC analysis, and those made for the most important alternative performance metrics such as the linear correlation coefficient and the root-mean-squared-difference (RMSD).

Moreover, different notations are being used to formulate and solve the TC problem, based either on cross-multiplied differences between the data sets, or on combinations of the (co-)variances

* Corresponding author.
E-mail address: alexander.gruber@geo.tuwien.ac.at (A. Gruber).

http://dx.doi.org/10.1016/j.jag.2015.09.002
0303-2434/

between them (Stoffelen, 1998; Loew and Schlenz, 2011; Scipal et al., 2008; Dorigo et al., 2010; Su et al., 2014b; McColl et al., 2014). This has fostered the impression of structurally different implementations, yet all proposed notations are mathematically identical. This identity will be analytically clarified in this study.

While the fundamental underlying maths and the required assumptions have remained unchanged over time, useful advances have been made in the way the obtained error estimates are presented and interpreted. In the literature, most studies investigate error variance estimates directly. Recently, several studies proposed to investigate errors relative to the underlying signal, i.e., as a direct or indirect representation of the signal-to-noise ratio (SNR) (Draper et al., 2013; Su et al., 2014b; McColl et al., 2014). Even less common than the investigation of the SNR is the investigation of soil moisture sensitivities, which can also be estimated using TC analysis (Stoffelen, 1998; McColl et al., 2014). In this study we will review the proposed metrics and demonstrate their similarities as well as their respective advantages and disadvantages. Finally, we propose the combined investigation of the SNR (expressed in logarithmic units), the unscaled error variances, and the soil moisture sensitivities of the data sets as an optimal combination to evaluate remotely sensed soil moisture data sets, which best exploits the complementary information content of the available performance metrics.

Section 2 compares the different notations used to formulate the TC problem. Section 3 provides a comprehensive discussion on the underlying assumptions. Section 4 compares different error presentations and demonstrates the proposed optimal evaluation strategy.

2. Triple collocation formulation

2.1. Error model

The most commonly used error model for TC analysis has the following form:

$$i = \alpha_i + \beta_i \Theta + \varepsilon_i \tag{1}$$

where $i \in [X, Y, Z]$ are three spatially and temporally collocated data sets. Θ is the unknown true soil moisture state; α_i and β_i are systematic additive and multiplicative biases of data set i with respect to the true state, and ε_i represents additive zero-mean random noise. Note that the additive bias α_i represents an offset between the temporal mean of data set i and the true soil moisture mean. Therefore, relative differences between α coefficients of different data sets can be easily corrected for by matching their temporal mean. Relative correction of the β coefficients is less trivial and will be discussed in Section 2.3. The underlying assumptions for the error model in (1) are: (i) Linearity between the true soil moisture signal and the observations, (ii) signal and error stationarity, (iii) independency between the errors and the soil moisture signal (error orthogonality), and (iv) independency between the errors of X, Y and Z (zero error cross-correlation). A detailed discussion on these assumptions will be provided in Section 3.

In TC analysis, the mean squared random error of all three data sets (i.e., the respective error variance $\sigma^2_{\varepsilon_i} = \langle \varepsilon_i^2 \rangle$, where $\langle \cdot \rangle$ denotes the temporal average) are estimated individually. Unlike the conventional (root-)mean-square-difference, TC estimates the error variances independently from the errors in a chosen reference data set. The most common way to solve for the $\sigma^2_{\varepsilon_i}$ is – as proposed by Stoffelen (1998) – by cross-multiplying differences between the three a-priori rescaled data sets. Stoffelen (1998) also proposed an alternative formulation (for the estimation of $\sigma^2_{\varepsilon_i}$), which is based on combinations of the covariances between the data sets. Even though both approaches are mathematically identical, the latter has been used only in a small number of recent studies (Loew and Schlenz, 2011; Su et al., 2014b,a; McColl et al., 2014). For the remainder of this paper, the former approach will be denoted as difference notation and the latter as covariance notation.

It is worth noting that standard triple collocation analysis based on (1) is a form of instrumental variable (IV) regression and that the framework of IV may provide an opportunity for extending the analyses to include several more variables (>3 data sets) and polynomial models (Su et al., 2014a; Bowden and Turkington, 1990). An alternative form of IV implementation is to use time-lagged versions of a data set as a third variable. Under the condition of weakly auto-correlated errors in the lagged variable, such an IV analysis yields the same results as TC but without the need for three coincident data sets. This is invaluable in practice when sampled data are limited due to limited spatio-temporal coverages of measuring systems or non-stationarity issues. For a detailed discussion on the relation between TC and IV we refer the reader to Su et al. (2014a) as this is beyond the scope of this paper.

2.2. Difference notation

When using the difference notation (Stoffelen, 1998; Scipal et al., 2008; Dorigo et al., 2010), the data sets first have to be rescaled against an arbitrarily chosen reference data set (this will be X for the following example). Subsequently, error variances can be estimated by averaging the cross-multiplied differences between the three data sets:

$$\begin{aligned}
\sigma^2_{\varepsilon_X} &= \langle (X - Y^X)(X - Z^X) \rangle \\
\sigma^2_{\varepsilon_Y^X} &= \langle (Y^X - X)(Y^X - Z^X) \rangle \\
\sigma^2_{\varepsilon_Z^X} &= \langle (Z^X - X)(Z^X - Y^X) \rangle
\end{aligned} \tag{2}$$

where the superscript X denotes the scaling reference. A detailed derivation of (2) is provided in Appendix A.

Since (2) requires rescaled data as input, it also estimates the error variances within the data space of the chosen scaling reference. Any error in the rescaling of the data will in turn lead to errors in the estimated error variances. In particular, these will not converge to the actual error variances, if the estimates of the scaling parameters themselves do not converge to their actual values as the number of samples increases. In other words, these scaling parameters have to be inferred using a consistent estimator.

2.3. Consistent estimation of scaling parameters

In the literature, many different rescaling techniques (e.g., linear regression, standardization, normalization, and others) have been applied. However, the only method that provides consistent estimates of (linear) scaling parameters also in case of differing signal-to-noise ratios (SNR) is triple collocation (Stoffelen, 1998; Yilmaz and Crow, 2013). It can be regarded as a form of instrumental variable regression, where a third variable (for instance, Z) is used as an instrument to resolve the relationship between erroneous measurements of two variables (X and Y) (Su et al., 2014a). Similarly, Y can act as an instrument to resolving the X–Z relationship. The resultant consistent estimates of the scaling factors β_i in these relationships yield the following solutions:

$$\begin{aligned}
\beta^*_Y &= \frac{\beta_X}{\beta_Y} = \frac{\langle (X - \bar{X})(Z - \bar{Z}) \rangle}{\langle (Y - \bar{Y})(Z - \bar{Z}) \rangle} = \frac{\sigma_{XZ}}{\sigma_{YZ}} \\
\beta^*_Z &= \frac{\beta_X}{\beta_Z} = \frac{\langle (X - \bar{X})(Y - \bar{Y}) \rangle}{\langle (Z - \bar{Z})(Y - \bar{Y}) \rangle} = \frac{\sigma_{XY}}{\sigma_{ZY}}
\end{aligned} \tag{3}$$

The overbar denotes the mean value of the time series, and β^*_X and β^*_Z are the rescaling coefficients which match the underlying true