Improvements of satellite observations through data merging: status and challenges

Seokhyeon Kim[1], Runze Zhang[2], Ashish Sharma[1], and Venkataraman Lakshmi[2]

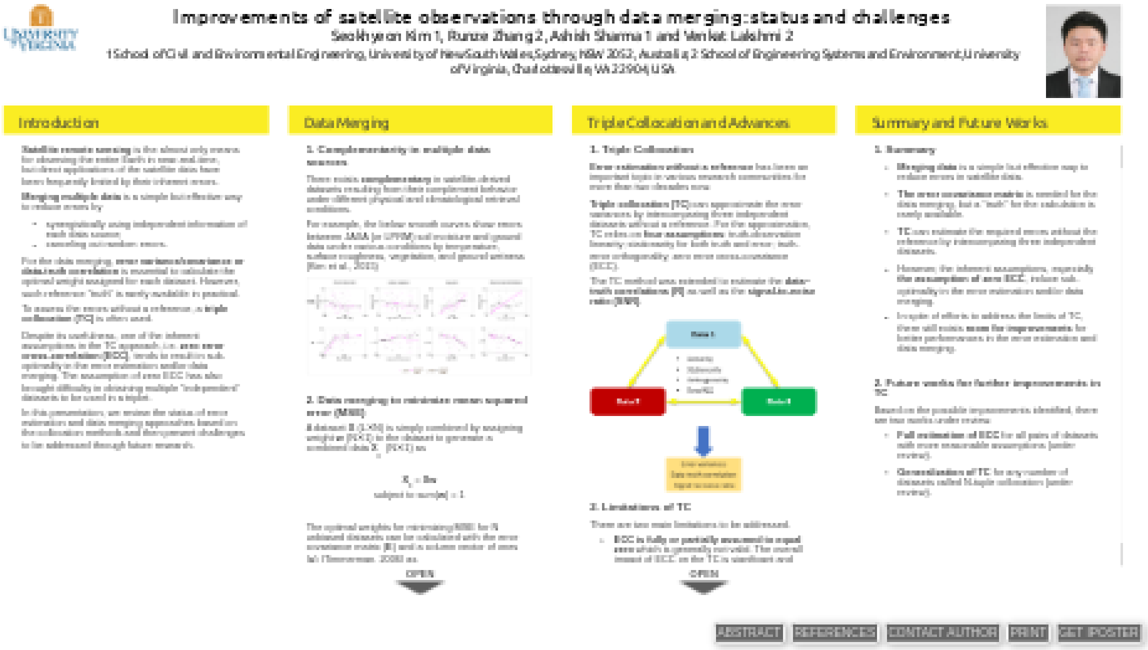
[1]University of New South Wales
[2]University of Virginia

November 22, 2022

Abstract

Satellite-derived data provide useful information about the rationale of Earth's functioning. While satellite remote sensing has been regarded as the almost only means for observing the entire Earth in near-real-time, errors in satellite observations have limited their direct usage in applications. Merging two or more data sources has been regarded as a simple but effective way to decrease such errors (e. g. minimizing mean square errors between the observation and truth). The principle of data merging is to combine independent information of each data source, improving over each individual product by canceling out random errors, with effectiveness by the degree of independence over the data sources. In the case of linearly combining data, qualitative assessments of the error (i.e. error variance/covariance and data-truth correlation) are essential to calculate the optimal weight for each candidate product. However, such reference "truth" is rarely available in practical. To overcome this limitation, a triple collocation (TC) technique is often used to estimate data error by using a data triplet without the truth. Despite the usefulness and simplicity of the TC-based error estimation, the inherent assumptions (e.g. error independence) in the approach tend to induce sub-optimal results in the error estimation and/or data combination. There have been also further efforts to address the limitation such as quadruple collocation (QC) using a data quadruple to partially estimate error cross-correlation and single/double instrumental variable methods to lessen the difficulty in obtaining multiple datasets. In this presentation, we review the status of error estimation and data merging approaches based on the collocation methods and then present challenges to be addressed through future research.

Improvements of satellite observations through data merging: status and challenges



Improvements of satellite observations through data merging:status and challenges

Seokhyeon Kim 1, Runze Zhang 2, Ashish Sharma 1 and Venkat Lakshmi 2

1 School of Civil and Environmental Engineering, University of New South Wales, Sydney, NSW 2052, Australia; 2 School of Engineering Systems and Environment, University of Virginia, Charlottesville, VA 22904, USA

Introduction

Satellite remote sensing is the almost only means for observing the entire Earth in near real-time, but direct applications of the satellite data have been frequently limited by their inherent errors.

Merging multiple data is a simple but effective way to reduce errors by

- synergistically using independent information of each data source;
- canceling out random errors.

For the data merging, **error variance/covariance or data-truth correlation** is essential to calculate the optimal weight assigned for each dataset. However, such reference "truth" is rarely available in practical.

To assess the errors without a reference, a **triple collocation** (TC) is often used.

Despite its usefulness, one of the inherent assumptions in the TC approach, i.e. **zero error cross-correlation (ECC)**, tends to result in sub-optimality in the error estimation and/or data merging. The assumption of zero ECC has also brought difficulty in obtaining multiple "independent" datasets to be used in a triplet.

In this presentation, we review the status of error estimation and data merging approaches based on the collocation methods and the present challenges to be addressed through future research.

Data Merging

1. Complementarity in multiple data sources

There exists **complementary** in satellite-derived datasets resulting from their complement behavior under different physical and climatological retrieval conditions.

For example, the below smooth curves show errors between JAXA (or LPRM) soil moisture and ground data under various conditions by temperature, surface roughness, vegetation, and ground wetness (Kim et al., 2015)

2. Data merging to minimize mean squared error (MSE)

A dataset X (LXN) is simply combined by assigning weights w (NX1) to the dataset to generate a combined data X_c (NX1) as

$X_c = Xw$

subject to sum(w) = 1

The optimal weights for minimizing MSE for N unbiased datasets can be calculated with the error covariance matrix (E) and a column vector of ones (u) (Timmermans, 2008) as

OPEN

Triple Collocation and Advances

1. Triple Collocation

Error estimation without a reference has been an important topic in various research communities for more than two decades now.

Triple collocation (TC) can approximate the error variances by intercomparing three independent datasets without a reference. For the approximation, TC relies on **four assumptions**: truth-observation linearity, stationarity for both truth and error, truth-error orthogonality, zero error cross-covariance (ECC).

The TC method was extended to estimate the **data-truth correlations (R)** as well as the **signal-to-noise ratio (SNR)**.

2. Limitations of TC

There are two main limitations to be addressed.

- **ECC is fully or partially assumed to equal zero** which is generally not valid. The overall impact of ECC on the TC is significant and

OPEN

Summary and Future Works

1. Summary

- **Merging data** is a simple but effective way to reduce errors in satellite data.
- **The error covariance matrix** is needed for the data merging, but a "truth" for the calculation is rarely available.
- **TC** can estimate the required errors without the reference by intercomparing three independent datasets.
- However, the inherent assumptions, especially **the assumption of zero ECC**, induce sub-optimality in the error estimation and/or data merging.
- In spite of efforts to address the limits of TC, there still exists **room for improvements** for better performances in the error estimation and data merging.

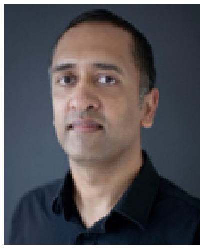
2. Future works for further improvements in TC

Based on the possible improvements identified, there are two works under review.

- **Full estimation of ECC** for all pairs of datasets with more reasonable assumptions (under review).
- **Generalization of TC** for any number of datasets called N-tuple collocation (under review).

Seokhyeon Kim 1, Runze Zhang 2, Ashish Sharma 1 and Venkat Lakshmi 2

1 School of Civil and Environmental Engineering, University of New South Wales, Sydney, NSW 2052, Australia; 2 School of Engineering Systems and Environment, University of Virginia, Charlottesville, VA 22904, USA



PRESENTED AT:

AGU FALL MEETING
Online Everywhere | 1–17 December 2020

INTRODUCTION

Satellite remote sensing is the almost only means for observing the entire Earth in near-real-time, but direct applications of the satellite data have been frequently limited by their inherent errors.

Merging multiple data is a simple but effective way to reduce errors by

- synergistically using independent information of each data source;
- canceling out random errors.

For the data merging, **error variance/covariance or data-truth correlation** is essential to calculate the optimal weight assigned for each dataset. However, such reference “truth” is rarely available in practical.

To assess the errors without a reference, a **triple collocation (TC)** is often used.

Despite its usefulness, one of the inherent assumptions in the TC approach, i.e. **zero error cross-correlation (ECC)**, tends to result in sub-optimality in the error estimation and/or data merging. The assumption of zero ECC has also brought difficulty in obtaining multiple "independent" datasets to be used in a triplet.

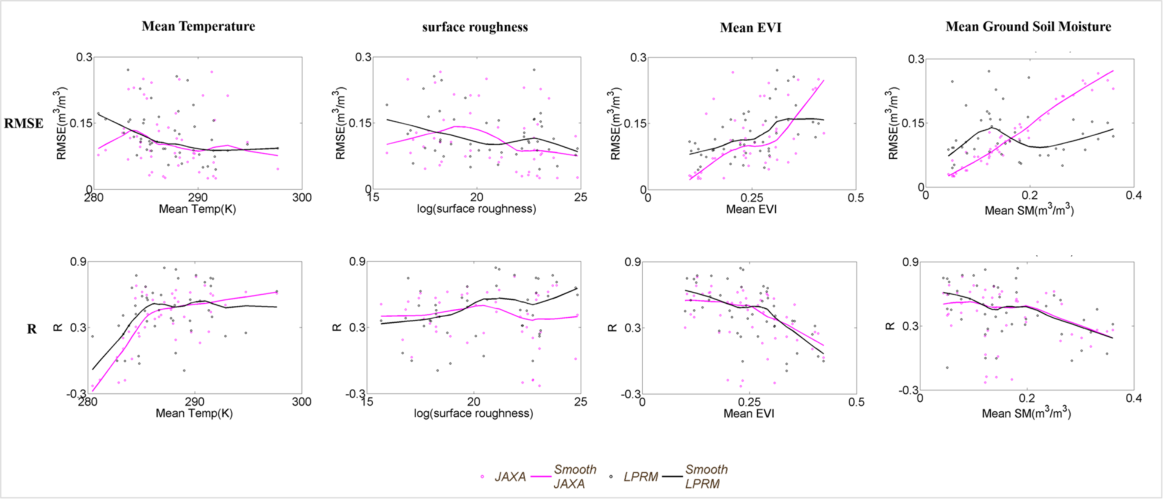
In this presentation, we review the status of error estimation and data merging approaches based on the collocation methods and then present challenges to be addressed through future research.

DATA MERGING

1. Complementarity in multiple data sources

There exists **complementary** in satellite-derived datasets resulting from their complement behavior under different physical and climatological retrieval conditions.

For example, the below smooth curves show errors between JAXA (or LPRM) soil moisture and ground data under various conditions by temperature, surface roughness, vegetation, and ground wetness (Kim et al., 2015)



2. Data merging to minimize mean squared error (MSE)

A dataset **X** (L×N) is simply combined by assigning weight **w** (N×1) to the dataset to generate a combined data $\mathbf{X}_c$ (N×1) as

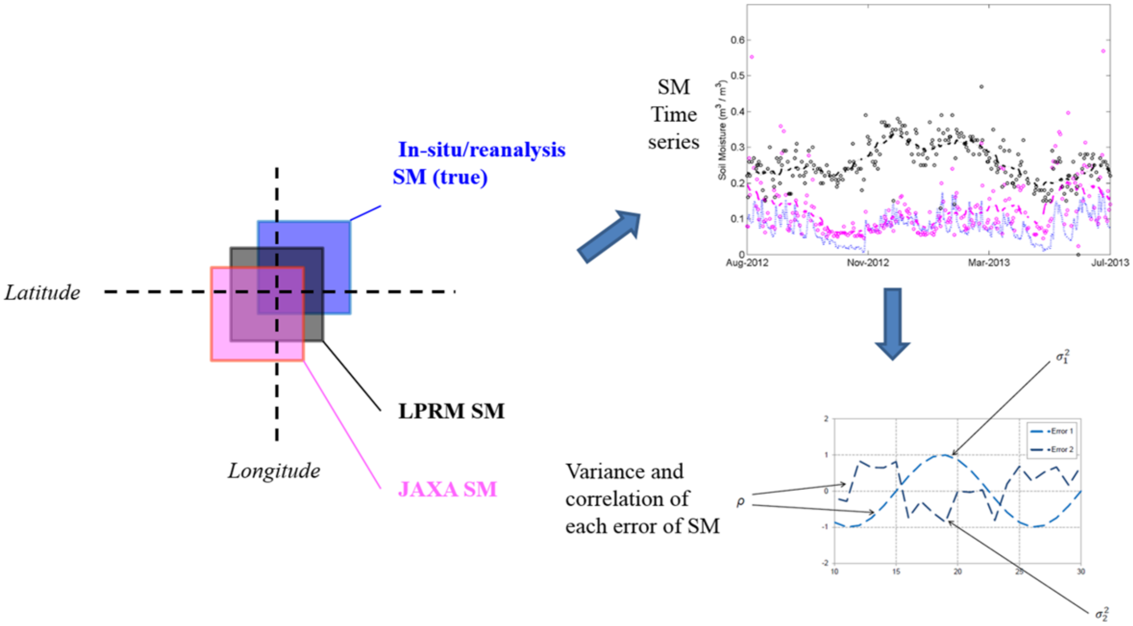
$$\mathbf{X}_c = \mathbf{X}\mathbf{w}$$

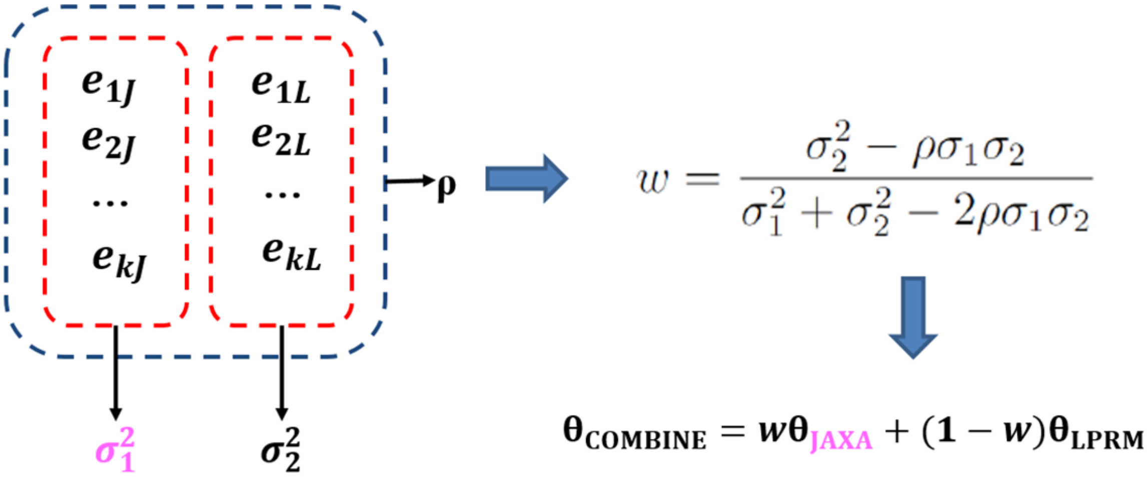
$$\text{subject to sum}(\mathbf{w}) = 1$$

The optimal weights for minimizing MSE for N unbiased datasets can be calculated with the error covariance matrix (**E**) and a column vector of ones (**u**) (Timmerman, 2006) as

$$\mathbf{w} = (\mathbf{u}^T\mathbf{E}^{-1}\mathbf{u})^{-1}\mathbf{E}^{-1}\mathbf{u}$$

For example, the following schematic diagrams show a case of merging two datasets.




e_{1J} e_{2J} … e_{kJ}
e_{1L} e_{2L} … e_{kL}
ρ
σ_1^2
σ_2^2
$w = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$
$\theta_{\mathrm{COMBINE}} = w\theta_{\mathrm{JAXA}} + (1 - w)\theta_{\mathrm{LPRM}}$

Here, the key information to calculate **w** is the error covariance matrix (**E**) which needs a reference "truth" that is seldom available.

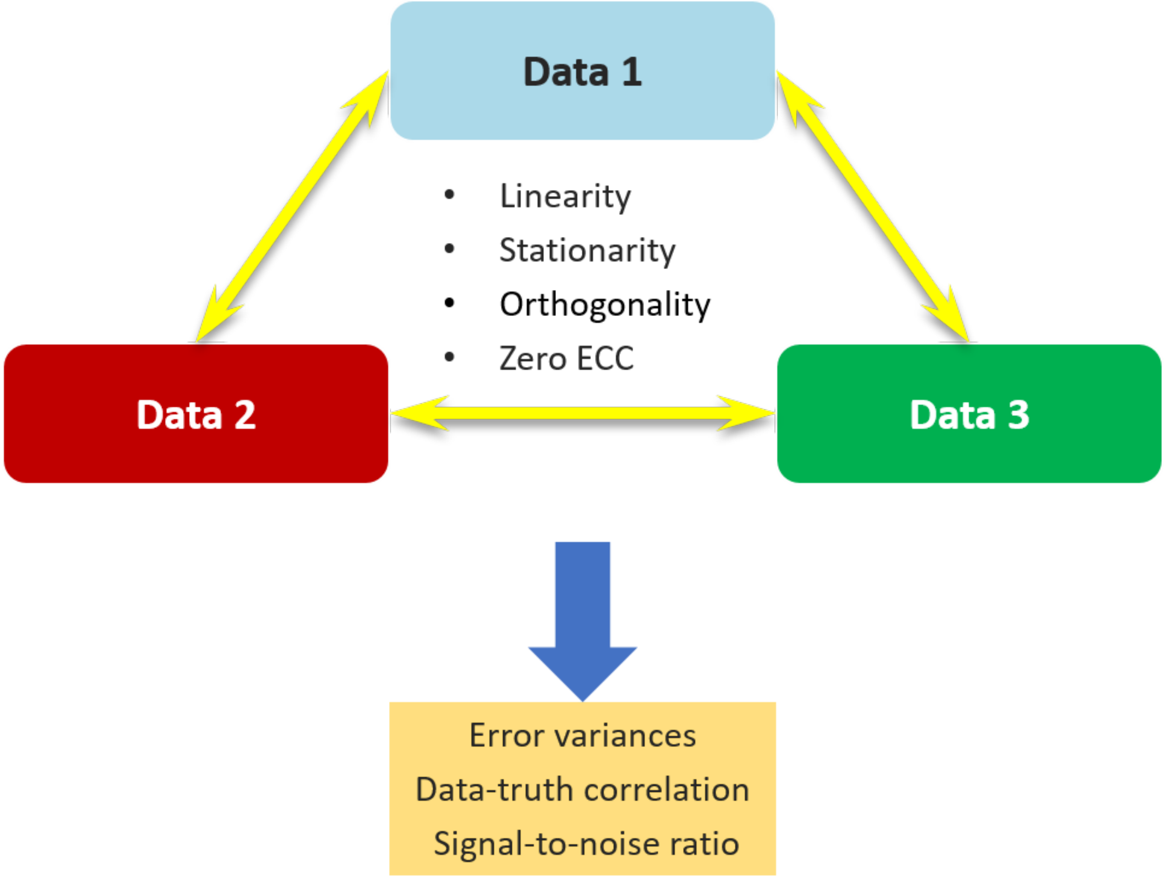
TRIPLE COLLOCATION AND ADVANCES

1. Triple Collocation

Error estimation without a reference has been an important topic in various research communities for more than two decades now.

Triple collocation (TC) can approximate the error variances by intercomparing three independent datasets without a reference. For the approximation, TC relies on **four assumptions**: truth-observation linearity; stationarity for both truth and error; truth-error orthogonality; zero error cross-covariance (ECC).

The TC method was extended to estimate the **data–truth correlations (R)** as well as the **signal-to-noise ratio (SNR)**.



2. Limitations of TC

There are two main limitations to be addressed.

- **ECC is fully or partially assumed to equal zero** which is generally not valid. The overall impact of ECC on the TC is significant and ignoring ECC can cause underestimation of errors. The assumption of zero ECC has also brought **difficulty in obtaining multiple "independent" datasets** to be used in a triplet.

- TC has not been generalized for any number of products.

3. Advances in TC

There has been a number of efforts to address the above-mentioned limits.

- **Quadruple collocation**: using a data quadruple to partially estimate the ECC

- **Single/double instrumental variable method**: using 1-day lag time series of the products as the third variable to lessen the difficulty in obtaining multiple datasets.

- **Extended double instrumental variable**: an improved version of the double instrumental variable method by which ECC can be partially estimated.

4. Room for improvements

In spite of such efforts, there still exists room for improvements in the existing collocation-based approaches, which should be addressed in future works.

- The ECC is to be fully estimated for all pairs of datasets to be estimated with more reasonable assumptions.

- It is useful for the method to be generalized for any number of datasets to be estimated (e.g. N-tuple collocation).

SUMMARY AND FUTURE WORKS

1. Summary

- **Merging data** is a simple but effective way to reduce errors in satellite data.
- **The error covariance matrix** is needed for the data merging, but a “truth” for the calculation is rarely available.
- **TC** can estimate the required errors without the reference by intercomparing three independent datasets.
- However, the inherent assumptions, especially **the assumption of zero ECC**, induce sub-optimality in the error estimation and/or data merging.
- In spite of efforts to address the limits of TC, there still exists **room for improvements** for better performances in the error estimation and data merging.

2. Future works for further improvements in TC

Based on the possible improvements identified, there are two works under review.

- **Full estimation of ECC** for all pairs of datasets with more reasonable assumptions (under review).
- **Generalization of TC** for any number of datasets called N-tuple collocation (under review).

ABSTRACT

Satellite-derived data provide useful information about the rationale of Earth’s functioning. While satellite remote sensing has been regarded as the almost only means for observing the entire Earth in near-real-time, errors in satellite observations have limited their direct usage in applications. Merging two or more data sources has been regarded as a simple but effective way to decrease such errors (e. g. minimizing mean square errors between the observation and truth). The principle of data merging is to combine independent information of each data source, improving over each individual product by canceling out random errors, with effectiveness by the degree of independence over the data sources. In the case of linearly combining data, qualitative assessments of the error (i.e. error variance/covariance and data-truth correlation) are essential to calculate the optimal weight for each candidate product. However, such reference “truth” is rarely available in practical. To overcome this limitation, a triple collocation (TC) technique is often used to estimate data error by using a data triplet without the truth. Despite the usefulness and simplicity of the TC-based error estimation, the inherent assumptions (e.g. error independence) in the approach tend to induce sub-optimal results in the error estimation and/or data combination. There have been also further efforts to address the limitation such as quadruple collocation (QC) using a data quadruple to partially estimate error cross-correlation and single/double instrumental variable methods to lessen the difficulty in obtaining multiple datasets. In this presentation, we review the status of error estimation and data merging approaches based on the collocation methods and then present challenges to be addressed through future research.

REFERENCES

Bates, J. M., & Granger, C. W. (1969). The combination of forecasts. Operational Research Quarterly, 20(4), 451-468.

Dong, J., Crow, W. T., Duan, Z., Wei, L., & Lu, Y. (2019). A double instrumental variable method for geophysical product error estimation. Remote Sensing of Environment, 225, 217-228.

Gruber, A., Su, C.-H., Crow, W. T., Zwieback, S., Dorigo, W. A., & Wagner, W. (2016). Estimating error cross-correlations in soil moisture data sets using extended collocation analysis. Journal of Geophysical Research: Atmospheres, 121(3), 1208-1219.

Hagan, D. F. T., Wang, G., Kim, S., Parinussa, R. M., Liu, Y., Ullah, W., . . . Su, B. (2020). Maximizing Temporal Correlations in Long-Term Global Satellite Soil Moisture Data-Merging. Remote Sensing, 12(13), 2164.

Kim, S., Liu, Y. Y., Johnson, F. M., Parinussa, R. M., & Sharma, A. (2015). A global comparison of alternate AMSR2 soil moisture products: Why do they differ? Remote Sensing of Environment, 161(0), 43-62.

Kim, S., Parinussa, R. M., Liu, Y. Y., Johnson, F. M., & Sharma, A. (2015). A framework for combining multiple soil moisture retrievals based on maximizing temporal correlation. Geophysical Research Letters, 42(16), 6662-6670.

Kim, S., Parinussa, R. M., Liu, Y. Y., Johnson, F. M., & Sharma, A. (2016). Merging Alternate Remotely-Sensed Soil Moisture Retrievals Using a Non-Static Model Combination Approach. Remote Sensing, 8(6), 518.

Kim, S., Pham, H., Liu, Y. Y., Marshall, L., & Sharma, A. (2020). Improving the combination of satellite soil moisture datasets by considering error cross-correlation: A comparison between triple collocation (TC) and extended double instrumental variable (EIVD) alternatives. IEEE transactions on Geoscience and remote sensing.

Kim, S., Pham, H., & Sharma, A. (Under review). A generalization of N-tuple collocation of multiple global soil moisture products without need for a reference.

McColl, K. A., Vogelzang, J., Konings, A. G., Entekhabi, D., Piles, M., & Stoffelen, A. (2014). Extended triple collocation: Estimating errors and correlation coefficients with respect to an unknown target. Geophysical Research Letters, 41(17), 6229-6236.

Su, C.-H., Ryu, D., Crow, W. T., & Western, A. W. (2014). Beyond triple collocation: Applications to soil moisture monitoring. Journal of Geophysical Research: Atmospheres, 119(11), 2013JD021043.

Timmermann, A. (2006). Forecast combinations. Handbook of economic forecasting, 1, 135-196.

Zhang, R., Kim, S., Sharma, A., & Lakshmi, V. (2021). Identifying relative strengths of SMAP, SMOS-IC, and ASCAT to capture temporal variability. Remote Sensing of Environment, 252, 112126.