Die approbierte Originalversion dieser Dissertation ist in der Hauptbibliothek der Technischen Universität Wien aufgestellt und zugänglich.
http://www.ub.tuwien.ac.at

The approved original version of this thesis is available at the main library of the Vienna University of Technology.
http://www.ub.tuwien.ac.at/eng

VIENNA UNIVERSITY OF TECHNOLOGY
DEPARTMENT OF GEODESY
AND GEOINFORMATION
RESEARCH GROUPS
PHOTOGRAMMETRY & REMOTE SENSING

DISSERTATION

Estimation of error structures in remotely sensed soil moisture data sets

Ausgeführt zum Zwecke der Erlangung des akademischen Grades eines

Doktors der technischen Wissenschaften (Dr.techn.)

unter der Leitung von
Univ.Prof. Dipl.-Ing. Dr.techn. Wolfgang Wagner

und der Mitbetreuung von
MSc. Dr.rer.nat. Wouter Dorigo

E120.1
Department für Geodäsie und Geoinformation
Forschungsgruppe Fernerkundung

eingereicht and der der Technischen Universität Wien
Fakultät für Mathematik und Geoinformation

von

Dipl.-Ing. Alexander Gruber

Matrikelnummer: 0825683
Grenzgasse 14-18/42
2340 Mödling
Österreich

Wien, am 9. März 2016 . . . .

VIENNA UNIVERSITY OF TECHNOLOGY
DEPARTMENT OF GEODESY
AND GEOINFORMATION
RESEARCH GROUPS
PHOTOGRAMMETRY & REMOTE SENSING

DISSERTATION

Estimation of error structures in remotely sensed soil moisture data sets

A thesis submitted in fulfillment of the academic degree of

Doktor der technischen Wissenschaften (Dr.techn.)*

under the supervision of
Univ.Prof. Dipl.-Ing. Dr.techn. Wolfgang Wagner

and the co-supervision of
MSc. Dr.rer.nat. Wouter Dorigo

E120.1
Department of Geodesy and Geoinformation
Research Group Remote Sensing

Research conducted at TU Wien
Faculty of Mathematics and Geoinformation

by

Dipl.-Ing. Alexander Gruber

Matriculation number: 0825683
Grenzgasse 14-18/42
2340 Moedling
Austria

Wien, am 9. März 2016 ....

* comparable to the Doctor of Engineering Sciences

Supervisor/Reviewer: Prof. Dr. Wolfgang Wagner
Department of Geodesy and Geoinformation
TU Wien
Gusshausstraße 27-29, 1040, Vienna, Austria ....
E-Mail: Wolfgang.Wagner@geo.tuwien.ac.at

Reviewer: Prof. Dr. Alexander Löw
Department of Geography
University of Munich
Luisenstraße 37, 80333, Munich, Germany ....
E-Mail: Alexander.Loew@lmu.de

Co-supervisor: Dr. Wouter Dorigo
Department of Geodesy and Geoinformation
TU Wien
Gusshausstraße 27-29, 1040, Vienna, Austria ....
E-Mail: Wouter.Dorigo@geo.tuwien.ac.at

Parts of the present work have already been published in several peer-reviewed journal papers, which include:

Gruber, A., Crow, W., Dorigo, W., & Wagner, W. (2015). The potential of 2D Kalman filtering for soil moisture data assimilation. Remote Sensing of Environment, 171, 137-148.

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, 45, 200-211.

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.

The work has been conducted as part of the following research projects:

ESA's Climate Change Initiative (CCI) for soil moisture (Contract No. 4000104814/11/I-NB)

eartH2Observe (European Union's Seventh Frame-work Programme, Grant Agreement No. 603608).

Erklärung zur Verfassung der Arbeit
Author's Statement

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

I herby declare that I independently drafted this manuscript, that all sources and references are correctly cited, and that the respective parts of this manuscript - including tables, maps, and figures - which were included from other manuscripts or the internet either semantically or syntactically are made clearly eveident in the text and all respective sources are correctly cited.

Dipl.-Ing. Alexander Gruber
Grenzgasse 14-18/42
2340 Mödling
Österreich

....

Acknowledgments

I would like to thank all people who made this dissertation possible. Special thanks go to my supervisors Wolfgang Wagner and Wouter Dorigo for their great mentoring, to Wade Crow for his excellent supervision during my stay in the United States, to Chun-Hsu Su for his great supervision during my stay in Australia, to Simon Zwieback for his brilliant co-authorship, to Alexander Löw for valuable comments and for reviewing this thesis, to the entire GEO Remote Sensing Research Group for a unique work environment and extraordinary teamwork, and to my family and friends who supported me throughout the years.

Finally, I want to thank also all people I haven't mentioned personally for valuable discussions and team work on conferences, workshops, meetings or elsewhere, and last but not least all scientists who I haven't met personally but who paved the road for this thesis.

"If I have seen further it is by standing on the shoulders of Giants."

- Isaac Newton -

Kurzfassung

Bodenfeuchte ist einer der wichtigsten Treiber im globalen Wasserkreislauf. Globale Bodenfeuchtemessungen sind daher unerlässlich um hydrologische Phänomene im System Erde wie den Klimawandel, Vegetationswachstum und andere zu erforschen. Die Wichtigste Quelle solcher Daten sind satellitengestützte Mikrowellensysteme, allerdings unterliegen die damit gewonnenen Beobachtungen diversen Ungenauigkeiten. Die korrekte Interpretation und Nutzung solcher Daten erfordert daher ein umfassendes Verständnis und die Kenntnis um ihre Fehler.

Die sogenannte Triple Collocation (TC) Analyse ist eine Methode zur individuellen Schätzung der Signal- und Fehlervarianzen von drei Datensätzen mit untereinander unkorrelierten Fehlern, ohne dabei einen hochgenauen Referenzdatensatz zu benötigen. Sie ist daher eine der wichtigsten Methoden zur Schätzung von Fehlerstrukturen in satellitenbasierten Bodenfeuchtedaten. Das volle Potential der Methode ist jedoch noch nicht voll ausgeschöpft und nach wie vor Gegenstand aktueller Forschung. Allerdings basiert die Methode auf einigen Annahmen über die Struktur der zu Grunde liegenden Daten, deren Gültigkeit ebenfalls noch nicht ausreichend untersucht wurde.

Ziel dieser Arbeit ist die Weiterentwicklung der TC Methode für eine verbesserte und vollständigere Beschreibung der Fehlerstrukturen von satellitenbasierten Bodenfeuchtemessungen. Bestehende Implementierungen der Methode werden Begutachtet, die zugrunde liegenden Annahmen evaluiert und die Methode erweitert beziehungsweise generalisiert zum Zwecke einer objektiveren Schätzung der Datenqualität von Bodenfeuchteprodukten, sowie zur Schätzung von räumlichen Fehler Autokorrelationsstrukturen und Fehler Kreuzkorrelationsstrukturen.

Abstract

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. The most important sources for global soil moisture records are space borne microwave instruments. However, such satellite-derived soil moisture products 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 individual signal- and random error variances of three collocated data sets with mutually uncorrelated errors without relying on a high-quality reference data set. It has therefore evolved as one of the most important methods for estimating error structures in remotely sensed soil moisture data sets. Nevertheless, the exploitation of the full potential of the TC method is still subject to ongoing research. On the other hand, TC analysis is based on a variety of assumption on the structure of the underlying data sets whose validity hasn't been fully investigated yet.

This thesis further develops the TC method, aiming for an improved and more complete estimation of error structures in remotely sensed soil moisture data sets. Existing TC implementations are reviewed, assumptions underlying the method are evaluated, and novel generalizations and extensions to the method are proposed, which allow for a more objective interpretation of soil moisture data quality as well as for the additional estimation of spatial error auto-correlation and mutual error cross-correlation structures.

Contents

List of Acronyms **xix**

List of Figures **xxi**

1 Introduction **3**

1.1 Study objectives 5

1.2 Summary of the publications 6

1.2.1 Publication I: Recent advances in (soil moisture) triple collocation analysis 6

1.2.2 Publication II: The potential of 2D Kalman filtering for soil moisture data assimilation 6

1.2.3 Publication III: Estimating error cross-correlations in soil moisture data sets using extended collocation analysis 7

1.2.4 Author's contributions 8

1.3 Scientific impact 9

2 Recent advances in (soil moisture) triple collocation analysis **11**

2.1 Introduction 12

2.2 Triple collocation formulation 13

2.2.1 Error model 13

2.2.2 Difference notation 14

2.2.3 Consistent estimation of scaling parameters 15

2.2.4 Covariance notation 15

2.3 Discussion of underlying assumptions 17

2.3.1 Linearity 17

2.3.2 Stationarity 18

2.3.2.1 Stationarity of the signal 18

2.3.2.2 Stationarity of the (random) errors 19

2.3.3 Error orthogonality 20

2.3.4 Zero error cross-correlation 20

2.3.5 Representativeness 21

2.3.6 Relation of assumptions in TC analysis and other performance metrics . . 22

2.4 Error presentation 24

2.4.1 Absolute error variance 24

2.4.2 Relative error variance 24

2.4.3 Logarithmic signal-to-noise ratio . . . 25
2.4.4 Demonstration . . . 27
2.5 Summary and Conclusion . . . 30

Appendices **33**

2.A Analytical investigation of the difference and covariance notations in triple collocation analysis . . . 33
2.A.1 Difference notation . . . 33
2.A.2 Covariance notation . . . 34
2.A.3 Mathematical identity of difference and covariance notations . . . 36
2.B The impact of representativeness errors on TC error variance estimates . . . 37
2.B.1 Two-scale problem . . . 37
2.B.2 Three-scale problem . . . 38

3 The potential of 2D Kalman filtering for soil moisture data assimilation **41**
3.1 Introduction . . . 42
3.2 Background . . . 43
3.2.1 2D Kalman filter . . . 43
3.2.2 Parameter estimation . . . 44
3.3 Demonstration . . . 47
3.3.1 Synthetic experiment . . . 47
3.3.1.1 Estimation of spatial error auto-correlation . . . 48
3.3.1.2 Impact on soil moisture analysis accuracy . . . 48
3.3.2 Real data experiment . . . 51
3.3.2.1 Estimation of spatial error auto-correlation . . . 52
3.3.2.2 Impact on soil moisture analysis accuracy . . . 54
3.4 Analytical investigation of the Kalman gain behavior . . . 57
3.5 Summary and conclusion . . . 60

Appendices **63**

3.A Justification for using a simplified steady-state Kalman gain expression . . . 63

4 Estimating error cross-correlations in soil moisture data sets using extended collocation analysis **65**
4.1 Introduction . . . 66
4.2 Background . . . 68
4.2.1 Triple collocation . . . 68
4.2.2 Extended collocation problem . . . 69
4.2.3 Least-squares solution . . . 70

4.2.4 Resolvable error structures 71
4.3 Demonstration 71
4.3.1 Synthetic experiment 72
4.3.2 Real data experiment 72
4.3.2.1 EC analysis over the ISMN 74
4.3.2.2 Representativeness errors in EC analysis 77
4.3.2.3 Global EC analysis 77
4.4 Summary and outlook 79
Bibliography **81**

List of Acronyms

1D	One-dimensional
2D	Two-dimensional
AMSR-E	Advanced Microwave Scanning Radiometer - Earth Observing System
ARS	Agricultural Research Service
ASCAT	Advanced Scatterometer
ATLAS	Auto-Tuned Land Data Assimilation System
CCI	Climate Change Initiative
CDF	Cummulative Distribution Function
CONUS	Contiguous United States
dB	Decibel
DGG	Discrete Global Grid
EC	Extended Collocation
ECMWF	European Centre for Medium-Range Weather Forecasts
ERA	ECMWF Reanalysis
ESA	European Space Agency
fMSE	fractional Mean-Square-Error
FOV	Field Of View
fRMSE	fractional Root-Mean-Square-Error
GLDAS	Global Land Data Assimilation System
GMS	Geosynchronous Meteorological Satellite
GOES	Geosynchronous Operational Environmental Satellite
IR	Infrared
ISMN	International Soil Moisture Network
IV	Instrumental Variable
KF	Kalman Filter
L3	Level 3
LPRM	Land Parameter Retrieval Model
LR	Little River
LW	Little Washita

METEOSAT	Meteorological Satellite
METOP	Meteorological Operational Satellite
NOAA	National Oceanic and Atmospheric Administration
NSR	Noise-to-Signal Ratio
PR	Precipitation Radar
QC	Quadruple Collocation
RC	Raynolds Creek
RFI	Radio Frequency Interference
RMSD	Root-Mean-Square-Difference
RMSE	Root-Mean-Square-Error
SCAN	Soil Climate Analysis Network
SMAP	Soil Moisture Active Passive
SMOS	Soil Moisture and Ocean Salinity
SNR	Signal-To-Noise Ratio
SSF	Surface State Flag
TC	Triple Collocation
TMI	TRMM Microwave Imager
TMPA	TRMM Multi-satellite Precipitation Analysis
TRMM	Tropical Rainfall Measuring Mission
ubRMSD	unbiased Root-Mean-Square-Difference
US	United States
USCRN	U.S. Climate Reference Network
USDA	United States Department of Agriculture
UTC	Universial Time Coordinated
VIRS	Visible and Infrared Scanner
VOD	Vegetation Optical Depth
VUA	VU University Amsterdam
WARP	Water Retrieval Package
WG	Walnut Gulch

List of Figures

2.1 a) Transformation function between the SNR[dB] (y-axis), and R_i^2 and fMSE$_i$ (x-axis), and b) between the SNR[dB] (y-axis) and the SNR in linear units (x-axis). The length of the major and minor axis of the grey ellipses correspond to the change in the respective metrics (SNR[-], SNR[dB], R_i^2, and fMSE$_i$), if the TC-based SNR estimate was wrong by 15%. 26
2.2 SNR[dB] estimates (left) and R_i^2 estimates (right) for ASCAT (top) and AMSR-E (bottom) over the contiguous United States (CONUS). Results are only shown for areas where all three data sets achieve a significant positive correlation ($p < 0.05$). . 28
2.3 SNR[dB] estimates (top), soil moisture sensitivity estimates (middle), and (unscaled) error variance estimates (bottom) for ASCAT (left) and AMSR-E (right) over the contiguous United States (CONUS). Results are only shown for areas where all three data sets achieve a significant positive correlation ($p < 0.05$). 29

3.1 The median and the inter-quartile-range (IQR) of estimated spatial error auto-correlations (y-axis) for different true auto-correlation levels (x-axis). The whisker length is 1.5*IQR. The dashed line is the 1:1 line. 48
3.2 Improvement of Pearson correlation (versus synthetic truth) of a 2D Kalman filtering analysis with respect to the 1D filtering case ($\Delta\rho_p$) for different δQ and δR. C_Q levels are plotted on the x-axis and C_R levels on the y-axis. 49
3.3 Station locations for the SCAN (green) and USCRN (red) soil moisture networks. . 52
3.4 Spatial error auto-correlations (averaged over all cardinal directions) for soil moisture estimates derived from TRMM using (3.1) (top) and for ASCAT surface soil moisture retrievals (middle), and the error auto-correlation difference between them (bottom). Red circles mark the USDA ARS watershed locations. 53
3.5 Spatial error ratios (averaged over all cardinal directions) for soil moisture estimates derived from TRMM using (3.1) (top) and for ASCAT surface soil moisture retrievals (bottom). Red circles mark the USDA ARS watershed locations. 55
3.6 Pearson correlation (ρ_p) between the open-loop model run without assimilating ASCAT observation (OL), the 1D Kalman filtering (1D) and 2D Kalman filtering (2D) estimates when assimilating ASCAT, and ground-based surface soil moisture observation results for individual 0.25-degree pixels located within each USDA ARS watershed site. Omnidirectionally averaged Kalman gain weights applied to observations within neighboring pixels are shown in brackets. 56

3.7 Pearson correlation (ρ_p) between the soil moisture analysis and measurements acquired at SCAN and USCRN sites (left) for both the 1D filter case (x-axis) and the 2D filter case (y-axis), and a histogram of the average weights that are given to the neighboring pixels in the 2D filter case at the site locations (right). 56
3.1 Optimal Kalman gain weight (based on the simplified expression given in (3.1)) for observations in neighboring pixels as a function of δR (x-axis) and δQ (y-axis) for different C_Q and ΔC ((a)-(d)). 59
3.2 Optimal Kalman gain weights for neighboring pixels calculated using the simplified expression in (3.1). Plotted values reflect an average across all cardinal directions and red circles mark the USDA ARS watershed locations. 60
3.A.1 Scatterplot of optimal weights (for observations in neighboring pixels) derived using the simplified analytical expression in (3.1) vs. full weights calculated via the numerical iteration of the full Kalman filter for all 16245 synthetic data sets. ρ_p is the Pearson- and ρ_s the Spearman correlation coefficient. Vertical lines represent the 5% and 95% quantile (dotted), the upper and the lower quartile (dashed), and the median (solid) of full optimal Kalman gain weights calculated at the 228 SCAN and USCRN sites (Figure 3.7). 64

4.1 The median and the inter-quartile-range (IQR) of estimated error cross-correlations (y-axis) at different true cross-correlation levels (x-axis) and different error variance levels. Whiskers represent 1.5 times the IQR. The estimation bias and RMSE averaged over all samples are provided on the top left hand corner. The sample size is 750 days. 73
4.2 The median and inter-quartile-range (IQR) of error cross-correlation estimates for soil moisture retrievals obtained from: the AMSR-E C- and X-band channels, ASCAT and the AMSR-E C-band channel, and ASCAT and the AMSR-E X-band channel for both absolute values (left) and anomalies (right). Whiskers represent 1.5 times the IQR. The sample size is 283 stations. 75
4.3 Estimated cross-correlations between the errors in absolute soil moisture retrievals obtained from: ASCAT and the AMSR-E C-band channel (left), ASCAT and the AMSR-E X-band channel (middle), and the AMSR-E C- and X-band channels (right) for the Contiguous United States, Europe, and New South Wales (Australia). 76
4.4 Estimated cross-correlations between the errors in anomalies of ASCAT and AMSR-E C-band channel (left), ASCAT and AMSR-E X-band channel (middle), and both AMSR-E C- and X-band channels (right) for the Contiguous United States, Europe, and New South Wales (Australia). 76
4.5 Global error cross-correlation estimates for AMSR-E C- and X-band soil moisture retrievals. White shading indicates areas where estimates did not converge to a meaningful value. 78

"Uncertainty is an uncomfortable position.
But certainty is an absurd one."

- Voltaire -

Chapter 1

Introduction

The invention of space-borne remote sensing techniques in the late 20th century has heralded a new era for our understanding of the Earth system by enabling frequent global observations of the state and dynamics of biogeophysical parameters. Soil moisture is among these one of the most important drivers of the global hydrological cycle, influencing processes such as land-atmosphere energy fluxes, climate change, vegetation growth, and many others (*Legates et al.*, 2011).

To date, the most important sources for global soil moisture records are satellite-based microwave radar and radiometer instruments (*Liu et al.*, 2011b). However, soil moisture retrievals from these instruments are - as all measurements - subject to errors which originate not only from noise in the measurement system itself, but also from simplifications, violated assumptions and miscalibrations in the retrieval models that are used to extract soil moisture information from raw satellite measurements. Therefore, the exploitation of soil moisture data requires an in-depth understanding of their accuracy.

Traditionally, the accuracy assessment of any measurement system is based on relative inter-comparison with high-quality reference data acquired from high-accuracy instruments under well-controlled laboratory or field conditions. However, the estimation of errors in satellite observations, particularly from microwave instruments, is not as straight forward because they sample the Earth surface with measurements that integrate over several hundreds of square kilometers. Such large footprints are virtually impossible to sample in situ, especially on a global scale which would be required to characterize the quality of soil moisture retrievals taken over different land cover types and under varying climatic conditions.

Nevertheless, soil moisture is a large-scale phenomenon that attains a certain degree of temporal stability (*Vachaud et al.*, 1985) which means that measurements taken at single locations can be - in the temporal domain - more or less representative for soil moisture dynamics over larger areas. Therefore, in the past, many studies have attempted to built up in situ networks with a high spatial sensor density in order to generate reliable reference data for the validation of satellite-based soil moisture retrievals (*Jackson et al.*, 2010). However, such high-density networks are still scarce on a global scale (*Dorigo et al.*, 2011a) and most available in situ reference data are - in addition to their inherent measurement errors - still subject to representativeness errors which often even exceed current soil moisture product accuracy targets. (*Miralles et al.*, 2010; *Gruber et al.*, 2013).

A remedy to this issue is provided by the so-called triple collocation (TC) method, which was first proposed by *Stoffelen* (1998) for sea winds and later applied also to soil moisture (*Scipal et al.*, 2008; *Dorigo et al.*, 2010) and other geophysical variables (*Vogelzang et al.*, 2011; *Caires and Sterl*, 2003; *Roebeling et al.*, 2012; *Fang et al.*, 2012). TC changes the paradigm of relying on a relative inter-comparison with high-quality reference data and instead uses three (erroneous) data sets with mutually uncorrelated errors in the analysis in order to simultaneously separate the signal and error components of each individual data set. More specifically, TC analysis allows for the estimation of both absolute random error variances of the data sets and relative scaling coefficients between the data sets. Absolute systematic errors with respect to the truth, however, remain unknown but they are usually anyhow considered as being of minor importance since most applications require information on temporal soil moisture dynamics rather than on absolute soil moisture levels (*Crow and Van den Berg*, 2010).

Even though TC analysis has been recognized as one of the most important methods for estimating error structures in remotely sensed soil moisture data sets there are still several research gaps. First of all, TC analysis is based on a variety of assumptions. Even though their validity is questioned frequently (*Yilmaz and Crow*, 2014), they have hardly been investigated, not least due to the lack of sufficient data or appropriate methods to do so (*Gruber et al.*, 2016a). However, while these assumptions are often considered unique to TC analysis it also hasn't been investigated in how far they are relevant to conventional relative error metrics.

The most critical assumption is probably the assumption of mutually uncorrelated errors between data sets (*Yilmaz and Crow*, 2014). A violation of this assumption induces biases in the TC-based error estimates. On the other hand, even if the errors of different data sets may be estimated in a consistent manner by carefully selecting independent triplets, the subsequent simultaneous use of data sets with correlated errors in any statistically rigorous data assimilation or merging scheme requires an accurate parameterization of their error correlation magnitude (*Crow et al.*, 2015). However, no method is currently available to provide such estimates.

While the estimation of error cross-correlations remains an unresolved issue, TC analysis has been applied to estimate temporal error auto-correlation structures (*Zwieback et al.*, 2013). Following this method, TC analysis may also be used to estimate error auto-correlations in the spatial domain. Such approach hasn't been reported in literature yet; most likely because spatial error auto-correlations do not affect the consistency of TC-based error variance estimates. Nonetheless, the existence of spatial error auto-correlation potentially allows for laterally propagating soil moisture information across grid cells using two-dimensional (2D) data assimilation strategies.

Finally, several studies have reported the potential of extending TC analysis to more than three data sets (*Zwieback et al.*, 2012; *Su et al.*, 2014a; *Pierdicca et al.*, 2015), yet it remains unknown which additional information can be gained from doing so.

1.1 Study objectives

This study aims to fill some of the above described research gaps by further developing the TC method. The specific objectives of the three research articles upon which this study is based are:

- **I Review on the state-of-the art in TC analysis**
  - (a) Reviewing existing TC implementations
  - (b) Evaluating the assumptions underlying TC analysis and investigating their relation to other performance metrics
  - (c) Investigating the contribution of representativeness errors to TC-based error estimates
  - (d) Identifying/defining an optimal TC-based approach for assessing soil moisture data quality
- **II Estimation of spatial error auto-correlation structures**
  - (a) Extending TC analysis for estimating spatial error auto-correlations in soil moisture data sets
  - (b) Implementing a 2D Kalman filter that utilizes spatial error auto-correlation estimates for propagating soil moisture information in space
  - (c) Analytically investigating the sensitivity of such filter to non-zero spatial error auto-correlations
- **III Estimation of error cross-correlation structures**
  - (a) Generalizing the TC method to allow for the inclusion of an arbitrary number of data sets
  - (b) Utilizing gained degrees of freedom for estimating error cross-correlations between soil moisture data sets
  - (c) Implementing a least-squares solution to the collocation system of equations to increase the precision of the estimates

1.2 Summary of the publications

1.2.1 Publication I: Recent advances in (soil moisture) triple collocation analysis

The first publication (*Gruber et al.*, 2016a) is a review paper on existing triple collocation implementations. Different notations which have been used in the past to formulate and solve the TC problem - either based on data set covariances or cross-multiplied differences - are shown to be mathematically identical. However, TC implementations differ in the way the obtained error estimates are presented: usually as (scaled) absolute error variances. More recently, also signal-to-noise ratio (SNR) related metrics, i.e., metrics which relate the error variances to the variance of the underlying soil moisture signal, have been proposed as more objective quality indicator. In this publication we propose and demonstrate 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.

Moreover, this publication provides an evaluation of the assumptions underlying the triple collocation method, which 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 the data set triplet (zero error cross-correlation). While these assumptions are often considered to be unique to the TC method we show analytically that they are also implicitly made in the application of other conventional performance metrics such as the correlation coefficient or the Root-Mean-Square-Difference (RMSD). Moreover, while a variety of diagnostic studies indicate that some of these assumptions are not always met, only few TC modifications have been proposed to mitigate the impact of violations thereof. This publication identifies the assumption of zero error cross-correlation as potentially the most critical one, yet up to now no method was available for testing its validity. A potential remedy to this problem is provided in the third publication (*Gruber et al.*, 2016b), where we propose an extension to the TC method which allows to estimate error cross-correlations under certain conditions (Section 1.2.3 and Chapter 4).

Finally, this publication shows analytically the impact of representativeness errors on the individual TC-derived error variance estimates of the data sets when they differ in their spatial resolution.

1.2.2 Publication II: The potential of 2D Kalman filtering for soil moisture data assimilation

The second publication (*Gruber et al.*, 2015) aims for the estimation and exploitation of spatial error auto-correlation structures in soil moisture data sets. While temporal error auto-correlations in single data sets and mutual error cross-correlations between data sets mainly introduce biases in the validation or application of soil moisture products, non-zero error auto-correlations in modelled and/or remotely sensed data provide an opportunity for laterally propagating soil moisture information to neighboring modelling pixels by utilizing a 2D data assimilation strategy.

To date, most 2D land data assimilation systems have been based on relatively crude and approximate guesses on error correlation structures. In this publication we propose a novel method as an extension of triple collocation analysis for reliably estimating spatial error auto-correlations. The method is validated using synthetic data sets and applied to drive a 2D Kalman filter both in a synthetic experiment and in a real-data scenario. Results of the real-data experiment are validated using well-controlled in situ reference stations as well as modelled reference soil moisture fields.

The synthetic experiment shows that the method is able to recover true spatial error auto-correlation levels without a bias and with negligible RMSE. Moreover, by using these estimates, a significant skill gain of the 2D filter with respect to a one-dimensional (1D) filter (in terms of correlation with the synthetic truth) is achieved. However, while considerable spatial error auto-correlation also exists in the errors for all three real-data products, the inclusion of this information into the 2D assimilation system does not significantly improve the performance of the system relative to to 1D baseline. This result is explained via an analytical evaluation of the impact of spatial error auto-correlation on the steady-state Kalman gain, which reveals that 2D filtering requires the existence of large error auto-correlation differences (between the assimilation model and the assimilated observations) in order to enhance the analysis relative to a 1D filtering baseline. As a result, large error auto-correlations alone (in both the model or the observations) are not sufficient to motivate the application of a 2D land data assimilation system.

Moreover, the findings of this publication reveal that (commonly made) crude assumptions of spatial error statistics in a 2D system will at best maintain the performance of a 1D approach or - more likely - worsen the filter forecasts because an over- or underestimation of error auto-correlation difference can lead to an overestimation (in absolute terms) of the Kalman gain weight for neighboring pixels.

1.2.3 Publication III: Estimating error cross-correlations in soil moisture data sets using extended collocation analysis

In the third publication (*Gruber et al.*, 2016b), the TC method is generalized to allow for the inclusion of an arbitrary number of data sets - referred to as extended collocation (EC) analysis - which not only increases the precision of the error estimates of the individual input data sets but also allows to relax the assumption of zero error-cross correlation between some data set combinations, which further allows for the estimation of some non-zero error cross-correlations. The number of non-zero error cross-correlations that can be estimated is limited by the overall number of data sets used and by their underlying error cross-correlation structure: Each member of the data set pairs with assumed non-zero error cross correlation must also be a member of at least one data set triplet with fully independent errors. Remaining degrees of freedom are used to solve the collocation system of equations in a least-squares sense.

A synthetic experiment shows that EC analysis is able to reliably recover true error cross-correlation levels. Applied to real soil moisture retrievals from the Advanced Microwave Scanning Radiometer-EOS (AMSR-E) C-band and X-band observations together with Advanced Scatterometer (ASCAT) retrievals, modeled data from the Global Land Data Assimilation System (GLDAS)-Noah and in situ measurements drawn from the International Soil Moisture Network (ISMN), EC yields reasonable and strong non-zero error cross-correlations between the two AMSR-E products. Against expectation, non-zero error cross-correlations are also found between ASCAT and AMSR-E.

Even though only demonstrated for four and five data sets, the EC method proposed in this publication is readily applicable to an arbitrary number of data sets, which would facilitate the estimation of more non-zero error cross-correlation terms (e.g., when using 3 passive data sets such as SMAP, AMSR2, and SMOS together with 2 active data sets such as ASCAT onboard MetOp-A and MetOp-B). Therefore, it represents an important step towards a fully-parameterized error covariance matrix which is vital for any rigorous data assimilation framework or data merging scheme.

1.2.4 Author's contributions

My personal contributions to these publications are as follows:

- **Publication I:** Conduction of the literature review, analytical proof of the similarity between different TC implementations, analytical demonstration of the similarity between assumptions made in TC and other conventional performance metrics, analytical investigation of the impact of representativeness errors on TC error estimates, demonstration of the logarithmic SNR as ideal validation strategy, major part of the writing

- **Publication II:** Development of the method for estimating spatial error auto-correlations, implementation and application of the 2D Kalman filter, validation of the proposed method, analytical evaluation of the empirical findings, major part of the writing

- **Publication III:** Development and implementation of the extended collocation method, validation of the method, major part of the writing

1.3 Scientific impact

The key findings and major scientific contributions of this study which - to my best knowledge - haven't been published before elsewhere in peer-reviewed literature are summarized as the following:

I Existing TC implementations and solutions are shown to be mathematically identical

II The logarithmic SNR is demonstrated as ideal representation for soil moisture data quality, especially for the relative comparison amongst different products

III It is shown analytically that the assumptions on error orthogonality and zero-error cross correlation are not unique to TC analysis but also implicitly made in other conventional (covariance-based) performance metrics such as the correlation coefficient or the RMSD

IV The contribution of representativeness errors to error estimates gleaned from TC analysis is shown analytically

V TC analysis is generalized to allow for the inclusion of an arbitrary number of data sets solving the collocation system of equation in a least-squares sense

VI The proposed generalization of TC allows to estimate error cross-correlation structures between soil moisture data sets

VII An alternative extension of TC is proposed which allows for the estimation of spatial error auto-correlations in soil moisture data sets

VIII It is shown both empirically and analytically that 2D filtering approaches require the existence of large spatial error auto-correlation differences between the assimilation model and the observation data set rather than the mere existence of large spatial error auto-correlations in the data sets

The scientific publications on which this study is based are presented in the following chapters.

Chapter 2

Recent advances in (soil moisture) triple collocation analysis

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.

---

[*] This chapter is an edited version of: 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, 45, 200-211.

2.1 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.*, 2011b), 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 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 2.2 compares the different notations used to formulate the TC problem. Section 2.3 provides a comprehensive discussion on the underlying assumptions. Section 2.4 compares different error presentations and demonstrates the proposed optimal evaluation strategy.

2.2 Triple collocation formulation

2.2.1 Error model

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

$$i = \alpha_i + \beta_i \Theta + \varepsilon_i \tag{2.1}$$

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 2.2.3.

The underlying assumptions for the error model in (2.1) 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 2.3.

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 (2.1) 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.

2.2.2 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:

$$\begin{aligned}
\sigma^2_{\varepsilon_X} &= \langle (X - Y^X)(X - Z^X) \rangle \\
\sigma^2_{\varepsilon_Y^X} &= \langle (Y^X - X)(Y^X - Z^X) \rangle \\
\sigma^2_{\varepsilon_Z^X} &= \langle (Z^X - X)(Z^X - Y^X) \rangle
\end{aligned} \tag{2.2}$$

where the superscript X denotes the scaling reference. A detailed derivation of (2.2) is provided in 2.A.

Since (2.2) 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.

2.2.3 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:

$$\begin{aligned} \beta_Y^* &= \frac{\beta_X}{\beta_Y} = \frac{\langle (X-\overline{X})(Z-\overline{Z}) \rangle}{\langle (Y-\overline{Y})(Z-\overline{Z}) \rangle} = \frac{\sigma_{XZ}}{\sigma_{YZ}} \\ \beta_Z^* &= \frac{\beta_X}{\beta_Z} = \frac{\langle (X-\overline{X})(Y-\overline{Y}) \rangle}{\langle (Z-\overline{Z})(Y-\overline{Y}) \rangle} = \frac{\sigma_{XY}}{\sigma_{ZY}} \end{aligned} \tag{2.3}$$

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 soil moisture signal variances or, more precisely, the soil moisture sensitivities (i.e., the variances of $\beta_i\Theta$ in (2.1)) through:

$$Y^X = \beta_Y^*(Y-\overline{Y}) + \overline{X} \qquad\qquad Z^X = \beta_Y^*(Z-\overline{Z}) + \overline{X} \tag{2.4}$$

Note that the above described scaling parameters could also be used to convert the scaled error variances - obtained from (2.2) - back into their own data space, but usually they are kept in a common data space to allow for a meaningful inter-comparison (see Section 2.4).

2.2.4 Covariance notation

An alternative approach for deriving the error variance (and scaling parameter) estimates is the aforementioned covariance notation (*Stoffelen*, 1998; *Loew and Schlenz*, 2011; *Su et al.*, 2014b,a; *McColl et al.*, 2014). It utilizes the data set variances (σ_i^2) and covariances (σ_{ij}), which can be written as:

$$\begin{aligned} \sigma_i^2 &= \beta_i^2\sigma_\Theta^2 + \sigma_{\varepsilon_i}^2 \\ \sigma_{ij} &= \beta_i\beta_j\sigma_\Theta^2 \end{aligned} \tag{2.5}$$

with $i, j \in [X, Y, Z]$ and $i \neq j$. σ_Θ^2 is the variance of the true jointly observed soil moisture signal; the term $\beta_i^2 \sigma_\Theta^2$ can be interpreted as the sensitivity of data set i to variations in this true signal. That is, the higher β_i, the stronger is the response of measurement i to soil moisture changes. Using the covariance notation allows to solve for the unscaled error variances directly through:

$$\begin{aligned} \sigma_{\varepsilon_X}^2 &= \sigma_X^2 - \frac{\sigma_{XY}\sigma_{XZ}}{\sigma_{YZ}} \\ \sigma_{\varepsilon_Y}^2 &= \sigma_Y^2 - \frac{\sigma_{YX}\sigma_{YZ}}{\sigma_{XZ}} \\ \sigma_{\varepsilon_Z}^2 &= \sigma_Z^2 - \frac{\sigma_{ZX}\sigma_{ZY}}{\sigma_{XY}} \end{aligned} \tag{2.6}$$

Using (2.5) also allows to estimate the TC-based rescaling parameters directly, as it can be seen from (2.3). A detailed derivation of (2.5) and (2.6) is provided in 2.A.

In summary, both the difference and the covariance notation can be used to estimate random error variances as well as (linear) rescaling parameters. In the difference notation, error variances are estimated within a common (arbitrarily chosen) reference data space, having the possibility of converting them back using the a priori (mandatorily) estimated scaling parameters. The covariance notation, on the other hand, directly estimates unscaled error variances, which could be then scaled into a common data space using a posteriori (optionally) estimated scaling parameters. The choice of whether to use scaled or unscaled error variances depends on the application, yet the choice of the notation is trivial as they provide identical results (see 2.A).

The reason for using the covariance notation instead of the difference notation is that it provides also estimates of the sensitivity of the data sets to soil moisture changes - represented by the $\beta_i^2 \sigma_\Theta^2$ parameters in (2.5) - in addition to the error variance estimates, whereas the difference notation estimates only the latter. These soil moisture sensitivity estimates are obtained through:

$$\begin{aligned} \beta_X^2 \sigma_\Theta^2 &= \frac{\sigma_{XY}\sigma_{XZ}}{\sigma_{YZ}} \\ \beta_Y^2 \sigma_\Theta^2 &= \frac{\sigma_{YX}\sigma_{YZ}}{\sigma_{XZ}} \\ \beta_Z^2 \sigma_\Theta^2 &= \frac{\sigma_{ZX}\sigma_{ZY}}{\sigma_{XY}} \end{aligned} \tag{2.7}$$

The investigation of soil moisture sensitivities provides additional opportunities for the validation and inter-comparison of data sets, which have not been fully exploited yet and will be discussed in Section 2.4. Before that, the following section discusses the assumptions that are made in TC analysis and the impact of their possible violation and/or relaxation.

2.3 Discussion of underlying assumptions

The assumptions underlying the error model used in Section 2.2 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).

Many studies investigated the validity of these assumptions and proposed alternative strategies to circumvent or minimize the impact of possible violations. Yet, the originally proposed formulation of *Stoffelen* (1998) (as shown in Section 2.2), which is based upon these assumptions, is still the most commonly used and most robust implementation. Even though a variety of diagnostic studies indicate that some of the assumptions are not always met and would require an adaptation of the model, few studies have proposed viable alternatives. Moreover, no evidence has been provided that any of the proposed adaptations led to an enhanced accuracy or reliability of the error estimates.

The following sections will provide a detailed discussion of the key assumptions and strategies that have been proposed to mitigate a violation of them.

2.3.1 Linearity

The assumption of linearity between the signal and the errors determines the shape of the error model in (2.1), i.e., $i = \alpha_i + \beta_i \Theta + \varepsilon_i$. It assumes the presence of additive and multiplicative biases (α_i and β_i) as well as additive zero-mean random noise (ε_i), and only zeroth- and first-order relationships to soil moisture.

While the covariance notation implies such linear model by definition (see Section 2.2.4), several studies have attempted to apply a non-linear model to the difference notation by using non-linear rescaling techniques such as cumulative distribution function (CDF-) matching. However, common non-linear methods will fail in matching the underlying soil moisture signal unless the SNR of the data sets are equal (as do also common linear methods other than the triple collocation based rescaling described in Section 2.2.3) (*Drusch, Wood, and Gao*, 2005; *Yilmaz and Crow*, 2013). This, in turn, can bias TC error estimates as discussed in Section 2.2.2. Moreover, this bias due to suboptimal scaling parameter estimation in a non-linear model might even exceed the bias introduced via the application of a linear model to a non-linear system, provided that the apparent non-linearities are small.

For instance, an in situ study by *De Lannoy et al.* (2007) indicates such a (third degree) polynomial relationship between point-scale soil moisture with field areal average. Non-linear relationship is also found between in situ, satellite-retrieved and modelled soil moisture by *Su and Ryu* (2015).

However, other studies also suggest that soil moisture dynamics become progressively more linear at coarser scales (*Crow and Wood*, 2002).

One important source of non-linearities is related to the aforementioned assumption of signal stationarity, which will be discussed in the following section.

2.3.2 Stationarity

In TC analysis stationarity is usually assumed for both the soil moisture signal and the random errors, i.e., their mean values and variances are assumed to remain constant over time.

2.3.2.1 Stationarity of the signal

Soil moisture is very unlikely to be stationary as rainfall and temperature patterns show a distinct seasonal pattern in most regions of the world, which results in a distinct climatology in soil moisture records. Such violation of signal stationarity does not affect TC analysis per se, as it considers temporally collocated triplets, which differ from a hypothetical stationary mean and variance by the same magnitude.

A problem arises if the climatology of the three used data sets differ from one another. This can happen for two reasons. Firstly, the data sets are very unlikely to have exactly the same spatial support. Therefore, they could be affected by physical processes which might in fact have different seasonal patterns, such as the growing cycle of different vegetation types which will influence the underlying soil moisture regime. A more comprehensive discussion on the impact of varying spatial supports will be provided in Section 2.3.5. Secondly, the data sets can have a systematic error in capturing the seasonal pattern, represented as temporally varying α_i and β_i coefficients in (2.1). This too is not a problem in itself, provided that the (wrong) variations of α_i and β_i are the same for all data sets. However, this is very unlikely as such variations usually arise from different sources, e.g., from a different (imperfect) vegetation treatment in the retrieval of the data sets.

No matter what the source, differences between the climatologies of the data sets manifest as non-linearities between them. These non-linearities are of significant importance as they might occur at different time scales (*Drusch et al.*, 2005; *Su and Ryu*, 2015). Consequently, their correction would involve multi-scale rescaling (*Su and Ryu*, 2015). In particular, time series from individual data sets can be decomposed into variations occurring at different time scales, and linear inter-data relations - as those in (2.1) - have to be treated at individual time scales separately with different β values.

As an alternative, many studies attempt to tackle the root of the problem by individually removing the climatology of the data sets directly, that is, transforming the observations into the anomaly space (*Stoffelen*, 1998; *Miralles et al.*, 2010; *Crow et al.*, 2012b; *Draper et al.*, 2013). However, this requires a reliable estimation of the climatology, which is susceptible to estimation errors especially when the data are erroneous and/or short, and to the chosen length of the intervals over which temporal averages are taken.

2.3.2.2 Stationarity of the (random) errors

TC usually requires a large data sample (i.e., for satellite soil moisture a data set covering several years) for the estimated random error variance to converge to the true value with a sufficient precision (*Zwieback et al.*, 2012). This estimate represents the average random error variance of the entire period. Invoking error stationarity requires that the error variance remains constant throughout the considered years and, more importantly, between different seasons. A violation of this assumption does not harm the reliability of the estimated average random error variance per se, but it limits its representativeness for particular subsets of the considered time period. Therefore, a time-variant characterization of errors might be beneficial for a large variety of applications (*Crow et al.*, 2005). Agricultural applications, for instance, require a precise error estimation at key crop development points within the growing season. Consequently, an error estimate that is dominated by large off-season errors would lead to a wrong judgement of the quality of the considered data set.

Such non-stationary random errors are very likely related to imperfections in the treatment of seasonalities of contributing processes (such as vegetation growth) in the retrieval model of the data sets. An apparent non-stationarity in the (true) soil moisture signal should have no impact on the stationarity of the errors as this would violate the assumption of error orthogonality (see Section 2.3.3).

Recently, *Loew and Schlenz* (2011) proposed a dynamic TC approach to obtain continuous fortnightly TC error estimates by applying TC analysis within 30-day windows centered over all fortnightly periods, respectively. However, the very short time period considered in this approach leads to an extremely low sampling density and thus to very low precision estimates (*Zwieback et al.*, 2012). Note that such a window-based approach can potentially account for time-variant biases between the data sets due to different underlying climatologies, as discussed in the previous section. If, by contrast, such time-variant biases are not accounted for, the deviations between the different soil moisture data sets will tend to persist over time. They are thus closely related to temporal auto-correlations of the errors (*Zwieback et al.*, 2013). The latter will reduce the precision, but not the consistency of the estimated error variances (*Zwieback et al.*, 2013).

An alternative approach to deal with non-stationary is to estimate multi-annual window-based error variances for each day of the year (*Su et al.*, 2014a). However, this approach reduces the sampling density significantly as compared to a classical implementation, which could also reduce the precision of the estimates below a critical value given the rather short temporal overlap of to date available independent data set triplets (5 years with approximately 1-3 daily measurements). Therefore, most studies rely on annual error variance estimates based on a large sampling density rather than on less precise seasonal estimates whose sampling uncertainties might exceed their actual inter-annual variability.

2.3.3 Error orthogonality

Error orthogonality means that the errors are independent from the true soil moisture signal, i.e., $\langle\theta\varepsilon_i\rangle = \sigma_{\theta\varepsilon_i} = 0$. Even though this assumption is commonly made in TC analysis, relatively little is known about its validity.

The first investigation of this assumption was recently made by *Yilmaz and Crow* (2014) both analytically and numerically using four heavily equipped in situ sites. Results of this study suggest that the assumption on error orthogonality does not hold for typical surface soil moisture data sets, yet the impact of this violation is generally negligible as the bias in error variance estimates due to error non-orthogonality is dampened by the application of rescaling parameters or even compensated if the magnitude of non-orthogonality is approximately the same for all considered data sets. However, if more than one time series is non-orthogonal, the non-orthogonality problem implies also cross-correlated errors, which were found by *Yilmaz and Crow* (2014) to be of greater importance than non-orthogonality. Error cross-correlations will be discussed in the following section.

2.3.4 Zero error cross-correlation

The validity of the assumption of independent (random) errors depends not only on the absence of error non-orthogonality but also on the choice of the data sets. Combinations, which are commonly assumed to fulfill this requirement are any triplets consisting of (i) active microwave retrievals, (ii) passive microwave retrievals, (iii) in situ measurements, or (iv) land surface models, provided that neither of them is dependent on another member of the triplet (e.g., a model that assimilates the microwave retrievals) (*Scipal et al.*, 2008; *Dorigo et al.*, 2010; *Crow and Van den Berg*, 2010; *Draper et al.*, 2013).

However, by investigating a set of these four observation types both numerically and analytically, *Yilmaz and Crow* (2014) recently found that significant non-zero error cross-correlations exist even in some of the data set combinations which are are commonly assumed to lack them, i.e., between active and passive satellite-based data. Moreover, they found that error cross-correlations have a greater influence on the error variance estimates than non-orthogonality because they are - unlike non-orthogonality - not compensated when being of equal magnitude for all data sets. Note, however, that the study of *Yilmaz and Crow* (2014) is based on heavily equipped watershed sites, which are rarely available on a global scale. No reliable method for estimating error cross-correlations over larger areas has yet been proposed even though this would serve a large variety of applications, in particular data assimilation.

2.3.5 Representativeness

From the error model in (2.1) it can be seen that all data sets are assumed to represent exactly the same soil moisture state, which is very unlikely, given that the three data sets typically have a different spatial measurement support (including also the sampling depth) and different sampling intervals (especially for satellite-derived data). However, soil moisture shows a large degree of temporal stability due to its time-integrative nature and also because most of its hydrological drivers (e.g., precipitation and evapotranspiration) take place at very large scales (*Vachaud et al.*, 1985; *Wagner et al.*, 2008). Consequently, a large fraction of the differences between the sampled (true) soil moisture states will be of a systematic nature, represented by different α_i and β_i coefficients, which do not affect the error variance estimates of TC analysis as they can be accounted for via appropriate rescaling. In addition, some parts of the differences between the sampled (true) soil moisture states can have a non-systematic nature. These differences originate from processes that affect only one or two of the data sets and can lead to biases in the TC-based error variance estimates, referred to as representativeness errors.

There are two likely scenarios in soil moisture TC analysis where representativeness errors might occur. In the first scenario, TC is applied on one point-scale in situ data set together with two coarse-scale data sets that have a comparable spatial representativeness such as active- and passive satellite retrievals. While all processes that lead to soil moisture variations at the in situ site also affect the coarse-scale average, there might be soil moisture variations within the support of the coarse-scale data sets that do not take place at the site location (e.g., localized rainfall events). In this case, TC will penalize the in situ site for its missing ability to resolve coarse-scale soil moisture features while the error variance estimate for the coarse-scale data sets will remain unbiased.

The second scenario is when applying TC on three data sets with significantly differing spatial representativeness such as a combination of an situ site, a medium-scale land surface model, and a coarse-scale satellite data set. In this case, TC will penalize both the point-scale and the coarse-scale data set with representativeness errors (manifested in different ways), while the error variance estimate for the medium-scale data set will remain unbiased.

Both scenarios are demonstrated analytically in 2.B. Note that particularly the first scenario should be seen as an opportunity rather than as a problem, because (i) it allows to estimate the coarse-scale representativeness of in situ sites (*Gruber et al.*, 2013; *Miralles et al.*, 2010; *Crow et al.*, 2012b), and (ii) it allows to estimate error variance estimates of coarse-scale data sets independent from the representativeness of the used in situ site. Even the second scenario allows for the unbiased error variance estimation of at least one of the data sets (i.e., the intermediate-resolution data set).

2.3.6 Relation of assumptions in TC analysis and other performance metrics

While the above described assumptions are being heavily discussed in the literature and often described as a unique limitation for TC analysis, it is worth noting that such assumptions can be equally important for most other conventional performance metrics used for the evaluation of coarse-scale soil moisture records.

The validation of coarse-resolution data usually requires an a priori rescaling in order to account for the systematic differences between the data sets due to different spatial support, measurement depth, sampling intervals, etc., which give rise to the issues discussed in Sections 2.3.1, 2.3.2 and 2.3.5, and, consequently, requires the assumptions on linearity, stationarity, and equal representativeness. Moreover, most other conventional performance metrics are - just like triple collocation - based on data set variances and covariances. Therefore, they too require the assumptions discussed in Sections 2.3.3 and 2.3.4, i.e., error orthogonality and zero error-cross correlation.

The most important conventional performance metrics - whose relation to the above described assumptions shall be demonstrated here - are the Pearson correlation coefficient (R) and the Root-Mean-Square-Difference (RMSD) (*Entekhabi et al.*, 2010). They are defined as:

$$\begin{aligned} \mathrm{R}_{ij} &= \frac{\sigma_{ij}}{\sqrt{\sigma_i^2 \sigma_j^2}} \\ \mathrm{RMSD}_{ij} &= \langle (i-j)^2 \rangle^{\frac{1}{2}} \end{aligned} \tag{2.8}$$

with $i \neq j$. As the Pearson correlation coefficient is linear, it is appropriate to apply the error model in (2.1), which allows to rewrite R using the full variance and covariance definitions from (2.A.8):

$$\mathrm{R}_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_i^2 \sigma_j^2}} = \frac{\beta_i \beta_j \sigma_\Theta^2 + \beta_j \sigma_{\Theta\varepsilon_i} + \beta_i \sigma_{\Theta\varepsilon_j} + \sigma_{\varepsilon_i \varepsilon_j}}{\sqrt{(\beta_i^2 \sigma_\Theta^2 + 2\,\beta_i \sigma_{\Theta\varepsilon_i} + \sigma_{\varepsilon_i}^2)(\beta_j^2 \sigma_\Theta^2 + 2\,\beta_j \sigma_{\Theta\varepsilon_j} + \sigma_{\varepsilon_j}^2)}} \tag{2.9}$$

Obviously, both error orthogonality and error cross-correlation terms are present. Assuming those to vanish - as it is made in TC analysis - simplifies (2.9) to:

$$\mathrm{R}_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_i^2 \sigma_j^2}} = \frac{\beta_i \beta_j \sigma_\Theta^2}{\sqrt{(\beta_i^2 \sigma_\Theta^2 + \sigma_{\varepsilon_i}^2)(\beta_j^2 \sigma_\Theta^2 + \sigma_{\varepsilon_j}^2)}} = \frac{1}{\sqrt{(1+\mathrm{NSR}_i)(1+\mathrm{NSR}_j)}} \tag{2.10}$$

where $\mathrm{NSR}_i = \frac{\sigma_{\varepsilon_i}^2}{\beta_i^2 \sigma_\Theta^2}$ and $\mathrm{NSR}_j = \frac{\sigma_{\varepsilon_j}^2}{\beta_j^2 \sigma_\Theta^2}$ are the noise-to-signal ratios of the data sets. The "signal" of the NSR refers to the soil moisture sensitivity of the respective data set and not to the true soil moisture signal variance.

Note that (2.10) does not contradict the common interpretation of the correlation coefficient. R describes the ability of a data set to capture temporal soil moisture changes, while (2.10) allows for a meaningful interpretation of the sources of decorrelation, namely the noise level of the contributing data sets relative to their respective soil moisture sensitivity. However, (2.10) additionally reveals that the interpretation of R is conditional on the same assumptions as made in TC analysis.

The dependency of the RMSD on the assumptions discussed in Section 2.3 can be easily shown by decomposing it into a bias-in-mean, bias-in-variance, and a correlation-dependent component (*Gupta et al.*, 2009):

$$\mathrm{RMSD}_{ij} = \sqrt{(\bar{i} - \bar{j})^2 + (\sigma_i - \sigma_j)^2 + 2\sigma_i\sigma_j(1 - \mathrm{R}_{ij})} \tag{2.11}$$

The implicit dependency of the RMSD on R, which becomes apparent from (2.11), directly shows that it too depends on the assumptions described above.

In summary, even though error orthogonality and zero error cross-correlation (besides the other above described assumptions) are commonly understood to be unique limitations of a TC analysis, one should keep in mind that they equally affect other conventional performance metrics. Here we picked R and the RMSD for demonstration purposes as they are the most commonly used metrics for (coarse-scale) soil moisture validation. However, the same considerations can be made for other (co-)variance based metrics (e.g. the slope of the linear regression, which is commonly used to evaluate downscaling performance).

It should be mentioned that conventional performance metrics usually assume - in addition to the previously described triple collocation assumptions - that the chosen reference data set is free of (random) errors. However, observations are never perfect and obtaining a reliable reference in particular for spatially integrated measurements is virtually impossible. Even in situ measurements, which are often used for this purpose, might be highly accurate on a point scale - provided that a proper calibration has been performed, which is not always the case - but they are very likely to contain representativeness errors (see Section 2.3.5) and are therefore of limited reliability at coarser scales (*Gruber et al.*, 2013). There are in fact heavily equipped in situ watersheds (*Jackson et al.*, 2010) - also referred to as core sites or cal/val sites - which are often expected to be more representative also at coarse scales but they are very limited in number on a global scale and still rely on a near-perfect sensor calibration, which is not always given. Triple collocation analysis, on the other hand, assigns these sensor- and representativeness errors to the in situ data set itself (see 2.B) and can be therefore considered in general as a more robust error estimation method than conventional metrics, also when using in situ data.

2.4 Error presentation

While the fundamental underlying error model and assumptions made in TC analysis remained relatively unchanged, significant progress has been made recently in the presentation and interpretation of error estimates. These advances will be presented in the following sections.

2.4.1 Absolute error variance

In the past, most evaluation studies focussed exclusively on the absolute error variance estimates obtained from TC analysis. As mentioned in Section 2.2.3, a meaningful inter-comparison of these estimates required them to be rescaled to a common reference data space in order to account for different units and for different soil moisture sensitivities of the data sets. Even though the choice of the reference does not influence the relative error ranking, it will introduce a dependency of the error estimates on the soil moisture sensitivity of the scaling reference. This is particularly important when comparing spatial error patterns as they will contain the spatial climatology of the reference data set, i.e., its soil moisture sensitivity pattern (see Figure 2 of *Draper et al.* (2013)). Also, absolute noise levels alone provide only limited information about actual data set quality, as will be shown in the following section.

2.4.2 Relative error variance

In order to overcome the dependency of scaled error variance patterns on the spatial climatology of the chosen scaling reference, *Draper et al.* (2013) recently proposed to normalize the unscaled error variance estimates with the corresponding total data set variance. The resulting metric is referred to as fractional root-mean-squared-error (fRMSE). For consistency reasons we will drop the square root here and use the fractional mean-squared-error (fMSE):

$$\mathrm{fMSE}_i = \frac{\sigma^2_{\varepsilon_i}}{\sigma^2_i} = \frac{\sigma^2_{\varepsilon_i}}{\beta^2_i \sigma^2_\Theta + \sigma^2_{\varepsilon_i}} = \frac{1}{1 + \mathrm{SNR}_i} \tag{2.12}$$

This provides a normalized representation of the signal-to-noise ratio ($\mathrm{SNR}_i = \frac{\beta^2_i \sigma^2_\Theta}{\sigma^2_{\varepsilon_i}}$), which is scaled between 0 and 1 where 0 means that the observed soil moisture signal is free of noise, 1 means that no soil moisture signal is observed at all, and 0.5 means that the noise variance is just as high as the observed soil moisture signal variance.

Relating the noise estimates to the underlying soil moisture sensitivity individually does not only remove the dependency of error patterns on the spatial sensitivity pattern of the scaling reference, it also allows for a better quantitative assessment of actual data quality. If there is a strong soil moisture signal or if the data set is very sensitive to changes therein - higher error levels are usually tolerable. Conversely, in areas with a very low soil moisture variability such as desert regions or very low soil moisture sensitivity, even a very low error level might be critical for a certain application. This aspect of data quality can not be investigated from absolute errors alone (*Su et al.*, 2014b).

Remember that the "signal" of the SNR refers to $\beta_i^2 \sigma_\Theta^2$, i.e., to the soil moisture sensitivity of the data set. Therefore, a measurement system that is overly sensitive to soil moisture changes (having a multiplicative bias $\beta_i > 1$) could have a better SNR than an unbiased system (i.e., $\beta_i \approx 1$) if their absolute noise level is equal. Moreover, multiplicative biases are usually of minor concern for coarse-scale data sets as their spatial support is usually not well defined (or known). Hence, they are usually rescaled to fit the model dynamics of a certain application. Such rescaling, on the other hand, preserves the SNR even if the bias is perfectly corrected for.

More recently, *McColl et al.* (2014) proposed to use TC analysis to estimate the (linear) correlation coefficient of the individual data sets with the underlying true signal. This is done by exchanging the error variance $\sigma_{\varepsilon_i}^2$ in the numerator of (2.12) with the signal sensitivity $\beta_i^2 \sigma_\Theta^2$, as it was estimated from (2.7):

$$\mathrm{R}_i^2 = \frac{\beta_i^2 \sigma_\Theta^2}{\beta_i^2 \sigma_\Theta^2 + \sigma_{\varepsilon_i}^2} = \frac{SNR_i}{1 + SNR_i} = \frac{1}{1 + \mathrm{NSR}_i} \tag{2.13}$$

Note that for consistency reasons we express the correlation to the true signal R_i as coefficient of determination, i.e., as R_i^2. This too is a normalized representation of the SNR_i, scaled between 0 and 1. In fact, it is just the reverse of the fMSE_i ($\mathrm{R}_i^2 = 1 - \mathrm{fMSE}_i$). However, an important insight can be gained when comparing it to the conventional linear (Pearson) correlation coefficient R_{ij} in (2.10). While R_i^2 is solely determined by the noise and the soil moisture sensitivity of data set i, R_{ij} is determined by the noise and the sensitivity of both contributing data sets. Consequently, if considering one data set as validation reference, R_{ij} will depend on its spatial error pattern, just like the scaled absolute error estimates depend on the spatial sensitivity pattern of the chosen scaling reference. That is, the TC-derived absolute coefficient of determination R_i^2 - and also the fMSE_i - are in any case better indicators of the data sets capability to capture temporal soil moisture variations than the Pearson correlation coefficient, also not least because all three metrics are based on the same underlying assumptions. Given the above described identical information content of fMSE_i and R_i^2 we will limit discussions of these metrics for the remainder of this paper solely to the fMSE_i.

2.4.3 Logarithmic signal-to-noise ratio

One issue in the interpretation of the fMSE_i is its inherent non-linear $(1+x)^{-1}$ behavior, as it can be seen in (2.12) and (2.13). If the SNR changes by a factor of λ, then the magnitude by which the fMSE_i changes due to this factor depends on the absolute fMSE_i level. Consequently, also the impact of uncertainties in fMSE_i estimates depend on the absolute fMSE_i level. The highest impact of error

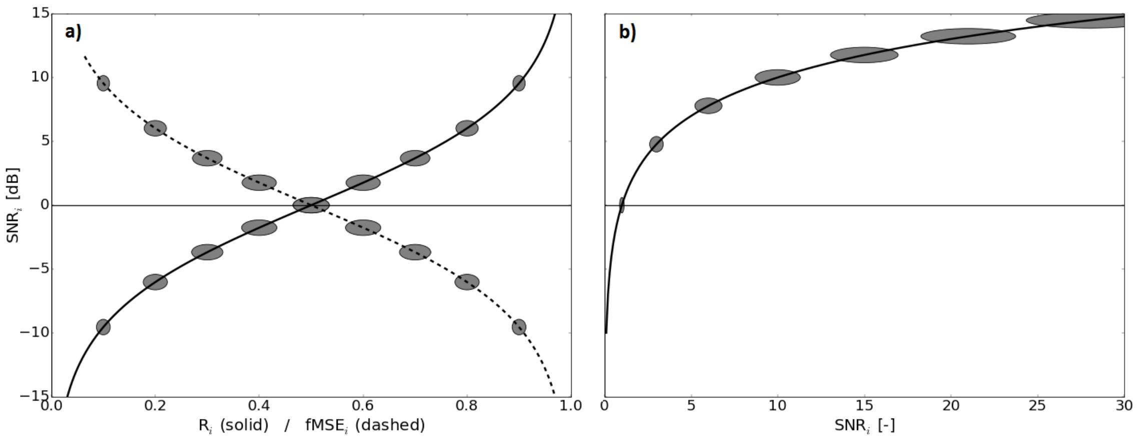

Figure 2.1: *a) Transformation function between the SNR[dB] (y-axis), and R_i^2 and fMSE$_i$ (x-axis), and b) between the SNR[dB] (y-axis) and the SNR in linear units (x-axis). The length of the major and minor axis of the grey ellipses correspond to the change in the respective metrics (SNR[-], SNR[dB], R_i^2, and fMSE$_i$), if the TC-based SNR estimate was wrong by 15%.*

or signal changes and, consequently, of estimation uncertainties on fMSE$_i$ estimates is found at fMSE$_i$ = 0.5, that is, if $\beta_i^2\sigma_\Theta^2 = \sigma_{\varepsilon_i}^2$.

For this reason, we propose to linearize the fMSE$_i$ behavior by using the SNR directly, and converting it into decibel units (dB):

$$\mathrm{SNR}_i[\mathrm{dB}] = 10\log(\mathrm{SNR}) = 10\log\left(\frac{\beta_i^2\sigma_\Theta^2}{\sigma_{\varepsilon_i}^2}\right) = -10\log\left(\frac{\sigma_i^2\sigma_{jk}}{\sigma_{ij}\sigma_{ik}} - 1\right) \tag{2.14}$$

where log(...) denotes the decadic logarithm. The use of the SNR as a performance metric, particularly when expressed in dB, is widespread in other disciplines such as electrical engineering and signal processing because of its convenient properties. Using dB makes the SNR symmetric around zero (see Figure 2.1). A value of zero means that the signal variance is equal to the noise variance, and every ± 3 dB correspond to an additional doubling/halving of the ratio between them. That is, +3(+6) dB means that the signal variance is twice (four times) as high as the noise variance, -3(-6) dB means that the signal variance is half (one fourth) of the noise variance, and so forth. Consequently, the SNR[dB] shows the highest sensitivity to performance changes in regions where $\beta_i^2\sigma_\Theta^2 \approx \sigma_{\varepsilon_i}^2$, which are usually of highest interest in cal/val studies because the largest performance improvements after algorithmic updates can be expected there. Conversely, extreme values of linear SNR estimates for very small $\beta_i^2\sigma_\Theta^2$ or $\sigma_{\varepsilon_i}^2$ are suppressed. Furthermore, due to the linearization also the impact of TC estimation uncertainties in (2.6) and (2.7) does not depend on the absolute SNR[dB] level any longer (as it is the case for the fMSE$_i$).

It should be emphasized that the proposed SNR[dB] does not provide different information than the fMSE$_i$; it simply eases a physically meaningful interpretation of the obtained numbers. This can be seen from Figure 2.1a, which shows the transformation function between them as well as between the SNR[dB] and the SNR in linear units. Note that in most parts of the important value ranges, i.e., between about +6 and -6 dB (where $\beta_i^2 \sigma_\Theta^2$ is between four times and one forth of $\sigma_{\varepsilon_i}^2$), the fMSE$_i$ and the SNR[dB] are approximately linearly related. The main difference is that the SNR[dB] is centered around zero, which may provide a better feeling for the physical meaning of the numbers. More important for the choice of using one or the other metric is the impact of estimation uncertainties, which is indicated as the major and minor axes of the ellipses in Figure 2.1a. One can see that TC estimation uncertainties have a greater impact on the fMSE$_i$ estimate, the closer the (linear) SNR is to 1. The SNR[dB], on the other hand, is equally sensitive to estimation uncertainties over its entire range. In Figure 2.1b, the need for the conversion to dB units becomes apparent, given the highly non-linear behavior of the SNR in linear units.

2.4.4 Demonstration

In this section we will demonstrate the strengths, weaknesses, and similarities of the metrics described above. Furthermore we will demonstrate the potential of investigating also soil moisture sensitivity estimates in addition to (unscaled) error variance and/or SNR estimates, which can provide useful complementary information that has not been fully exploited yet. Therefore, TC analysis is applied to soil moisture data acquired from: (i) active satellite retrievals, (ii) passive satellite retrievals, and (iii) a land surface model, covering the time period from January 2007 to October 2011.

The active satellite-based soil moisture data set is the H-25 SM-OBS-4 MetOp-A ASCAT time series product, retrieved using the TU Wien algorithm version WARP 5.5 R2.1 (*Wagner et al.*, 1999; *Naeimi*, 2009). ASCAT operates at C-band and estimates soil moisture as degree of saturation at a spatial resolution of 25 km, regridded to a 12.5 km Discrete Global Grid (DGG) and with a temporal resolution of 1-3 days. The WARP Surface State Flag (SSF; *Naeimi et al.*, 2012) was used to remove measurements taken under frozen or freezing/thawing conditions.

Passive satellite-based soil moisture estimates are retrieved from the Advanced Microwave Scanning Radiometer - Earth Observing System (AMSR-E) onboard the Aqua satellite, using the Land Parameter Retrieval Model (LPRM) Version 5 (*Owe et al.*, 2008), provided by the VU University Amsterdam (VUA). Both C-band and X-band observations are considered, which achieve a spatial resolution of 73*43 km^2 and 54 *30 km^2, respectively. Data is provided in volumetric units on a regular grid with 0.25 degrees grid spacing. The Vegetation Optical Depth (VOD) estimates are used to filter out retrievals with a high uncertainty due to dense vegetation, while Radio Frequency Interference (RFI) estimates are used to switch from C- to X-band retrievals in RFI-contaminated areas (*Owe et al.*, 2008).

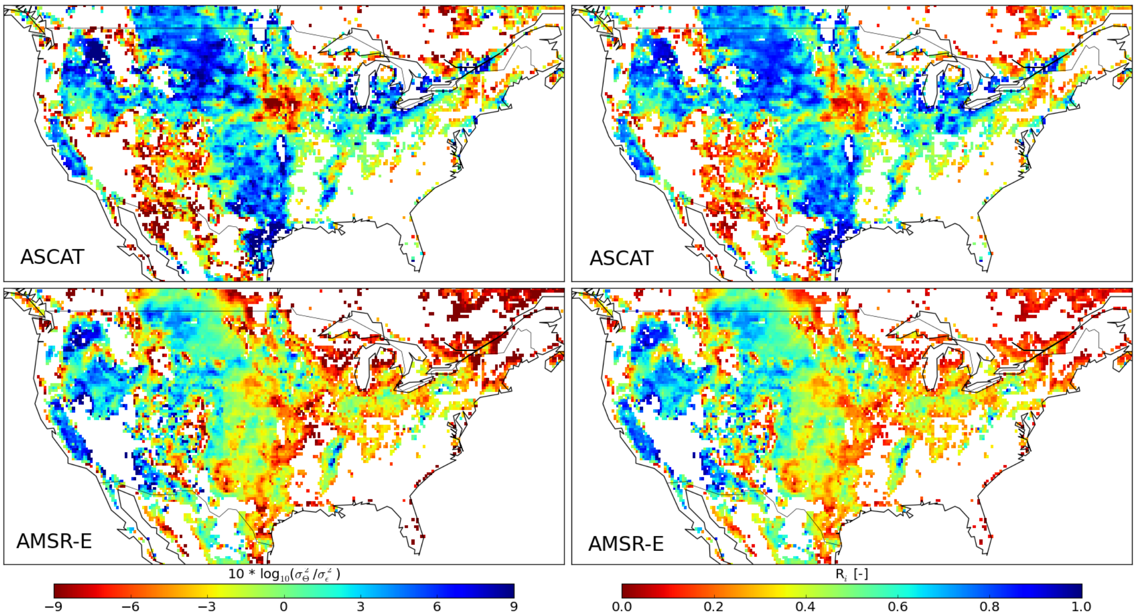

Figure 2.2: *SNR[dB] estimates (left) and R_i^2 estimates (right) for ASCAT (top) and AMSR-E (bottom) over the contiguous United States (CONUS). Results are only shown for areas where all three data sets achieve a significant positive correlation ($p < 0.05$).*

The land surface model that was used for this study is the Global Land Data Assimilation System (GLDAS-) Noah model, which provides soil moisture data for four different depth layers at a spatial resolution of approximately 0.25 degrees in a 3-hourly sampling rate. Only the top layer (0-10 cm) was used in this study. GLDAS-Noah also provides soil temperature estimates and an estimate of snow water equivalent. These were used to mask soil moisture measurements for which the temperature was below 0°C and for which the snow water equivalent estimate was not zero.

Figure 2.2 shows a comparison between the SNR[dB] and R_i^2 over the contiguous US (CONUS). The majority of pixels achieve a SNR[dB] between -9 and +9, which is the range where both metrics are approximately linearly related (see Figure 2.1). Slight, but negligible differences in the visual pattern can only be found at the positive and negative extremes. However, we prefer the SNR[dB] over the R_i^2 for the reasons discussed in Section 2.4.3. That is, because (i) the interpretation of the physical meaning of the numbers is more intuitive, and (ii) it is less sensitive to estimation uncertainties in these value ranges (see Figure 2.1). Note that the $fMSE_i$ pattern would look identical to that of the R_i^2 but with an inverted value range as $fMSE_i = 1 - R_i^2$. For a comparison between the SNR[dB] (or the $fMSE_i$) and scaled error variances we refer the reader to *Draper et al.* (2013) who comprehensively demonstrated the benefit of investigating SNR-related estimates rather than scaled absolute error variances.

Novel insights can be gained through the simultaneous evaluation of the SNR[dB] (or the $fMSE_i$) together with both unscaled error variances and the soil moisture sensitivities, which is shown in

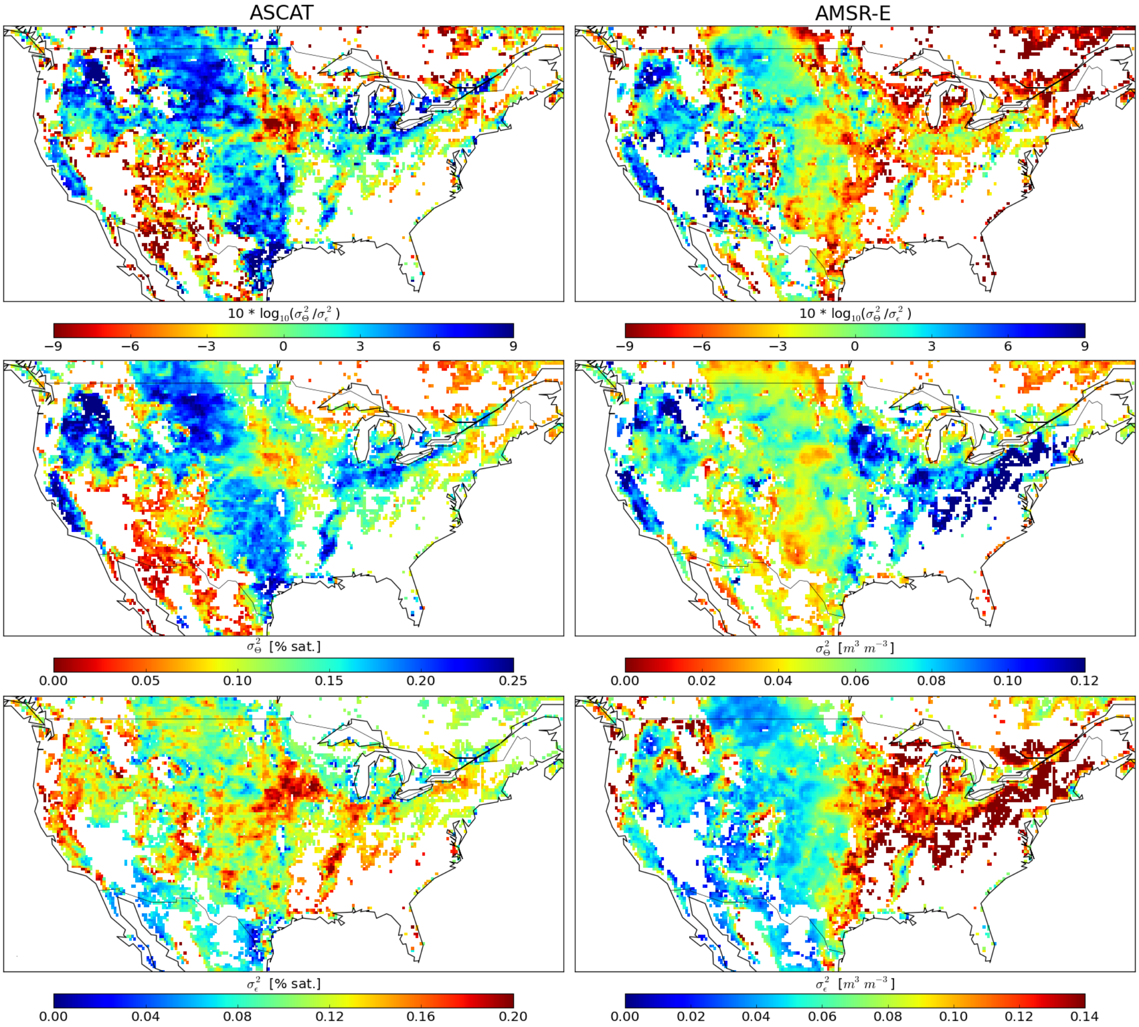
ASCAT
AMSR-E
10 * log$_{10}$($\sigma_\Theta^2 / \sigma_\epsilon^2$)
−9
−6
−3
0
3
6
9
σ_Θ^2 [% sat.]
0.00
0.05
0.10
0.15
0.20
0.25
σ_Θ^2 [$m^3\ m^{-3}$]
0.00
0.02
0.04
0.06
0.08
0.10
0.12
σ_ϵ^2 [% sat.]
0.00
0.04
0.08
0.12
0.16
0.20
σ_ϵ^2 [$m^3\ m^{-3}$]
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14

Figure 2.3: *SNR[dB] estimates (top), soil moisture sensitivity estimates (middle), and (unscaled) error variance estimates (bottom) for ASCAT (left) and AMSR-E (right) over the contiguous United States (CONUS). Results are only shown for areas where all three data sets achieve a significant positive correlation ($p < 0.05$).*

Figure 2.3. The SNR[dB] pattern of ASCAT pretty much coincides with its sensitivity pattern, while the error variances are rather homogeneous over CONUS (except for some areas with slightly larger error variances, which are concentrated mainly in the central US). On the contrary, the signal sensitivity of AMSR-E shows a well-pronounced west-east gradient (low sensitivity in the west and high sensitivity in the east) while its SNR[dB] shows an inverse behavior. This is because it is dominated by a strongly pronounced negative west-east gradient of the error variances. In other words, the SNR[dB] of ASCAT is mainly dominated by its sensitivity pattern whereas the SNR[dB] of AMSR-E is mainly dominated by its error variance pattern.

Such findings may have an important impact on the development and improvement of novel and existing retrieval models as they allow to pinpoint areas in which the sensor and/or the algorithm are prone to noise and in which areas they have a reduced sensitivity to soil moisture. Therefore, such areas can be related to geographic features such as rainfall patterns or vegetation, as targeting these problems requires different strategies. However, this is beyond the scope of this paper.

2.5 Summary and Conclusion

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. It aims to estimate the random error variances of three collocated data sets and does not require the availability of a high-quality reference data set.

Different notations have been used in the past to formulate and solve the TC problem - based either on cross-multiplied differences or on covariances between the data sets - which has fostered the impression of structurally different implementations. In this study, the mathematic identity of existing notations was demonstrated analytically.

Furthermore, a detailed discussion of the assumptions underlying TC analysis was provided. Even though a variety of diagnostic studies indicate that some of the assumptions are not always met, few studies have proposed viable alternatives to mitigate violations of the assumptions. Moreover, we demonstrated that most of the assumptions made in TC analysis, which are often considered to be unique to the method, are also implicitly made in the application of other conventional performance metrics.

However, the number of - particularly satellite based - measurement systems for the global observation of soil moisture dynamics is rapidly increasing, not only at coarse scales (e.g., through the recent launch of the Soil Moisture Active Passive (SMAP) mission) but also at higher spatial resolution (e.g., through the recently launched Sentinel-1 mission). Moreover, also the temporal coverage of existing soil moisture records is consistently growing. This progress can be expected to facilitate the development of advanced strategies to tackle issues related to violations of assumptions made in TC analysis, in particular issues related to non-stationarity, non-linearity, and error cross-correlation.

Useful advances have been made in the way the obtained error estimates are presented and interpreted, particularly using signal-to-noise ratio (SNR) metrics, i.e., by relating the error variances to the variance of the underlying soil moisture signal. A comparison of existing metrics and a discussion on their similarity as well as their respective advantages and disadvantages was provided.

Finally, we proposed 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 evaluation combination for remotely sensed soil moisture data sets, which best exploits the complementary information content of the available performance metrics. This will not only help in the understanding and improvement of existing satellite- or model-based soil moisture data sets, but will also facilitate the development of new models and retrieval algorithms for novel soil moisture missions such as SMAP.

Appendix

2.A Analytical investigation of the difference and covariance notations in triple collocation analysis

As mentioned in Section 2.2 are the difference and the covariance notations mathematically identical formulations of the same problem. This section will show the derivation of the target estimates for both notations in detail in order to demonstrate their identity analytically. Starting point is again the error model which was introduced in Section 2.2.1:

$$\begin{aligned} X &= \alpha_X + \beta_X \Theta + \varepsilon_X \\ Y &= \alpha_Y + \beta_Y \Theta + \varepsilon_Y \\ Z &= \alpha_Z + \beta_Z \Theta + \varepsilon_Z \end{aligned} \tag{2.A.1}$$

2.A.1 Difference notation

For the difference notation one data set is arbitrarily chosen as reference data set (this will be data set X for the following demonstration), against which the other two data sets are rescaled in a linear fashion:

$$Y^X = \beta_Y^*(Y - \overline{Y}) + \overline{X} \qquad\qquad Z^X = \beta_Z^*(Z - \overline{Z}) + \overline{X} \tag{2.A.2}$$

where $\beta_Y^* = \frac{\beta_X}{\beta_Y}$ and $\beta_Z^* = \frac{\beta_X}{\beta_Z}$. The superscript denotes the scaling reference. 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, matching their temporal mean matches also their α coefficients. The multiplicative rescaling parameters β_Y^* and β_Z^* can be derived from (2.A.1) by combining the three data sets in the following way:

$$\begin{aligned} \frac{\langle (X - \overline{X})(Z - \overline{Z}) \rangle}{\langle (Y - \overline{Y})(Z - \overline{Z}) \rangle} &= \frac{\beta_X \beta_Z \langle (\Theta - \overline{\Theta})^2 \rangle + \beta_X \langle (\Theta - \overline{\Theta}) \varepsilon_Z \rangle + \beta_Z \langle (\Theta - \overline{\Theta}) \varepsilon_X \rangle + \langle \varepsilon_X \varepsilon_Z \rangle}{\beta_Y \beta_Z \langle (\Theta - \overline{\Theta})^2 \rangle + \beta_Y \langle (\Theta - \overline{\Theta}) \varepsilon_Z \rangle + \beta_Z \langle (\Theta - \overline{\Theta}) \varepsilon_Y \rangle + \langle \varepsilon_Y \varepsilon_Z \rangle} \\ \frac{\langle (X - \overline{X})(Y - \overline{Y}) \rangle}{\langle (Z - \overline{Z})(Y - \overline{Y}) \rangle} &= \frac{\beta_X \beta_Y \langle (\Theta - \overline{\Theta})^2 \rangle + \beta_X \langle (\Theta - \overline{\Theta}) \varepsilon_Y \rangle + \beta_Y \langle (\Theta - \overline{\Theta}) \varepsilon_X \rangle + \langle \varepsilon_X \varepsilon_Y \rangle}{\beta_Z \beta_Y \langle (\Theta - \overline{\Theta})^2 \rangle + \beta_Z \langle (\Theta - \overline{\Theta}) \varepsilon_Y \rangle + \beta_Y \langle (\Theta - \overline{\Theta}) \varepsilon_Z \rangle + \langle \varepsilon_Z \varepsilon_Y \rangle} \end{aligned} \tag{2.A.3}$$

Since the errors are assumed to have zero mean, $\langle \varepsilon_X \varepsilon_X \rangle = \sigma^2_{\varepsilon_X}$, $\langle \varepsilon_Y \varepsilon_Y \rangle = \sigma^2_{\varepsilon_Y}$, and $\langle \varepsilon_Z \varepsilon_Z \rangle = \sigma^2_{\varepsilon_Z}$ represent the error variances of X, Y, and Z, respectively, $\langle \varepsilon_X \varepsilon_Y \rangle = \sigma_{\varepsilon_X \varepsilon_Y}$, $\langle \varepsilon_X \varepsilon_Z \rangle = \sigma_{\varepsilon_X \varepsilon_Z}$, and $\langle \varepsilon_Y \varepsilon_Z \rangle = \sigma_{\varepsilon_Y \varepsilon_Z}$ represent the error covariances between X, Y, and Z, respectively, and $\langle (\Theta - \overline{\Theta}) \varepsilon_X \rangle =$

$\sigma_{\Theta\varepsilon_X}$, $\langle(\Theta - \overline{\Theta})\varepsilon_X\rangle = \sigma_{\Theta\varepsilon_Y}$, and $\langle(\Theta - \overline{\Theta})\varepsilon_Z\rangle = \sigma_{\Theta\varepsilon_X}$ represent the covariances between the soil moisture signal and the errors of X, Y, and Z, respectively. As described in Section 2.3, the errors are assumed to be uncorrelated ($\sigma_{\varepsilon_X\varepsilon_Y} = 0$, $\sigma_{\varepsilon_X\varepsilon_Z} = 0$, and $\sigma_{\varepsilon_Y\varepsilon_Z} = 0$) and orthogonal ($\sigma_{\Theta\varepsilon_X} = 0$, $\sigma_{\Theta\varepsilon_Y} = 0$, and $\sigma_{\Theta\varepsilon_Z} = 0$). Therefore, (2.A.3) reduces to a consistent estimate of the rescaling parameters:

$$\begin{aligned} \beta_Y^* &= \frac{\beta_X}{\beta_Y} = \frac{\langle(X - \overline{X})(Z - \overline{Z})\rangle}{\langle(Y - \overline{Y})(Z - \overline{Z})\rangle} = \frac{\sigma_{XZ}}{\sigma_{YZ}} \\ \beta_Z^* &= \frac{\beta_X}{\beta_Z} = \frac{\langle(X - \overline{X})(Y - \overline{Y})\rangle}{\langle(Z - \overline{Z})(Y - \overline{Y})\rangle} = \frac{\sigma_{XY}}{\sigma_{ZY}} \end{aligned} \tag{2.A.4}$$

The rescaled data sets can then be written as:

$$\begin{aligned} X &= \alpha_X + \beta_X\Theta + \varepsilon_X \\ Y^X &= \alpha_X + \beta_X\Theta + \beta_Y^*\varepsilon_Y \\ Z^X &= \alpha_X + \beta_X\Theta + \beta_Z^*\varepsilon_Z \end{aligned} \tag{2.A.5}$$

Averaging over cross-multiplied differences of these rescaled data sets yields:

$$\begin{aligned} \langle(X - Y^X)(X - Z^X)\rangle &= \langle\varepsilon_X\varepsilon_X\rangle - \beta_Y^*\langle\varepsilon_X\varepsilon_Y\rangle - \beta_Z^*\langle\varepsilon_X\varepsilon_Z\rangle + \beta_Y^*\beta_Z^*\langle\varepsilon_Y\varepsilon_Z\rangle \\ \langle(Y^X - X)(Y^X - Z^X)\rangle &= {\beta_Y^*}^2\langle\varepsilon_Y\varepsilon_Y\rangle - \beta_Y^*\langle\varepsilon_Y\varepsilon_X\rangle - \beta_Y^*\beta_Z^*\langle\varepsilon_Y\varepsilon_Z\rangle + \beta_Z^*\langle\varepsilon_X\varepsilon_Z\rangle \\ \langle(Z^X - X)(Z^X - Y^X)\rangle &= {\beta_Z^*}^2\langle\varepsilon_Z\varepsilon_Z\rangle - \beta_Z^*\langle\varepsilon_Z\varepsilon_X\rangle - \beta_Z^*\beta_Y^*\langle\varepsilon_Z\varepsilon_Y\rangle + \beta_Y^*\langle\varepsilon_X\varepsilon_Y\rangle \end{aligned} \tag{2.A.6}$$

The above described assumption of zero error cross-correlation reduces also (2.A.6) to a consistent estimator for the (rescaled) error variances:

$$\begin{aligned} \sigma_{\varepsilon_X}^2 &= \langle(X - Y^X)(X - Z^X)\rangle \\ {\beta_Y^*}^2\sigma_{\varepsilon_Y}^2 = \sigma_{\varepsilon_Y^X}^2 &= \langle(Y^X - X)(Y^X - Z^X)\rangle \\ {\beta_Z^*}^2\sigma_{\varepsilon_Z}^2 = \sigma_{\varepsilon_Z^X}^2 &= \langle(Z^X - X)(Z^X - Y^X)\rangle \end{aligned} \tag{2.A.7}$$

Note that even though the error variances of Y and Z are estimated within the data space of the chosen scaling reference (X in this case), they could be converted back into their own data space a posteriori since the rescaling coefficients β_Y^* and β_Z^* are known from (2.A.4).

2.A.2 Covariance notation

For the covariance notation, the error model (2.A.1) can be considered as a sum of two random variables, namely soil moisture (Θ) and the random error (ε_i) with $i \in [X, Y, Z]$. Consequently, the (co-)variances of the data sets can be written as:

$$\begin{aligned} \sigma_i^2 &= \beta_i^2 \sigma_\Theta^2 + 2\,\beta_i \sigma_{\Theta\varepsilon_i} + \sigma_{\varepsilon_i}^2 \\ \sigma_{ij} &= \beta_i \beta_j \sigma_\Theta^2 + \beta_j \sigma_{\Theta\varepsilon_i} + \beta_i \sigma_{\Theta\varepsilon_j} + \sigma_{\varepsilon_i\varepsilon_j} \end{aligned} \tag{2.A.8}$$

with $i, j \in [X, Y, Z]$. If we further assume error orthogonality ($\sigma_{\Theta\varepsilon_i} = 0$) and zero error-cross correlation ($\sigma_{\varepsilon_i\varepsilon_j} = 0$ for $i \neq j$), (2.A.8) simplifies to:

$$\begin{aligned} \sigma_i^2 &= \beta_i^2 \sigma_\Theta^2 + \sigma_{\varepsilon_i}^2 \\ \sigma_{ij} &= \beta_i \beta_j \sigma_\Theta^2 \end{aligned} \tag{2.A.9}$$

The variance of data set i is thus a sum of its error variance ($\sigma_{\varepsilon_i}^2$) and the term $\beta_i^2 \sigma_\Theta^2$ (i.e., the true soil moisture variance σ_Θ^2 multiplied with the (squared) multiplicative bias of the data set), which reflects the sensitivity of data set i to soil moisture changes. That is, the higher β_i, the stronger is the response of i to soil moisture changes. This soil moisture sensitivity can be estimated for each data set individually by combining their covariances in the form:

$$\begin{aligned} \beta_X^2 \sigma_\Theta^2 &= \frac{\sigma_{XY}\sigma_{XZ}}{\sigma_{YZ}} \\ \beta_Y^2 \sigma_\Theta^2 &= \frac{\sigma_{YX}\sigma_{YZ}}{\sigma_{XZ}} \\ \beta_Z^2 \sigma_\Theta^2 &= \frac{\sigma_{ZX}\sigma_{ZY}}{\sigma_{XY}} \end{aligned} \tag{2.A.10}$$

Estimates of the error variances $\sigma_{\varepsilon_i}^2$ can be then be obtained by subtracting the soil moisture sensitivities of the data sets from their total variance:

$$\begin{aligned} \sigma_{\varepsilon_X}^2 &= \sigma_X^2 - \frac{\sigma_{XY}\sigma_{XZ}}{\sigma_{YZ}} \\ \sigma_{\varepsilon_Y}^2 &= \sigma_Y^2 - \frac{\sigma_{YX}\sigma_{YZ}}{\sigma_{XZ}} \\ \sigma_{\varepsilon_Z}^2 &= \sigma_Z^2 - \frac{\sigma_{ZX}\sigma_{ZY}}{\sigma_{XY}} \end{aligned} \tag{2.A.11}$$

Note that the error variances are now - unlike in the difference notation - estimated in their own data space. However, even though the covariance notation does not require an a priori rescaling of the data sets, (2.A.9) also allows for the direct estimation of the linear rescaling parameters:

$$\beta_Y^* = \frac{\beta_X}{\beta_Y} = \frac{\sigma_{XZ}}{\sigma_{YZ}} \qquad \beta_Z^* = \frac{\beta_X}{\beta_Z} = \frac{\sigma_{XY}}{\sigma_{ZY}} \tag{2.A.12}$$

which is identical to (2.A.4). These rescaling parameters could be used to convert the error variance estimates obtained from (2.A.11) into a common reference data space a posteriori, which would yield values identical to those obtained from (2.A.7), i.e., to the error estimates which were estimated in the reference data space directly using the difference notation. This mathematical identity will be demonstrated in the following section.

2.A.3 Mathematical identity of difference and covariance notations

The mathematical identity of the difference and the covariance notation can be shown by decomposing (2.A.7) (i.e., the error variance estimates obtained using the difference notation) similar to the decomposition of the RMSD, which was made by *Gupta et al.* (2009). Let us therefore insert (2.A.2) into (2.A.7) and extract the mean value of the unscaled data set:

$$\begin{aligned} \sigma^2_{\varepsilon_X} &= \langle[\overline{X} + (X - \overline{X}) - \beta^*_Y(Y - \overline{Y}) + \overline{X}][\overline{X} + (X - \overline{X}) - \beta^*_Z(Z - \overline{Z}) + \overline{X}]\rangle \\ {\beta^*_Y}^2\sigma^2_{\varepsilon_Y} &= \langle[\beta^*_Y(Y - \overline{Y}) + \overline{X} - (\overline{X} + (X - \overline{X}))][\beta^*_Y(Y - \overline{Y}) + \overline{X} - \beta^*_Z(Z - \overline{Z}) + \overline{X}]\rangle \\ {\beta^*_Z}^2\sigma^2_{\varepsilon_Z} &= \langle[\beta^*_Z(Z - \overline{Z}) + \overline{X} - (\overline{X} + (X - \overline{X}))][\beta^*_Z(Z - \overline{Z}) + \overline{X} - \beta^*_Y(Y - \overline{Y}) + \overline{X}]\rangle \end{aligned} \tag{2.A.13}$$

Expanding the brackets and completing the average yields:

$$\begin{aligned} \sigma^2_{\varepsilon_X} &= \sigma^2_X - \beta^*_Y\sigma_{XY} - \beta^*_Z\sigma_{XZ} + \beta^*_Y\beta^*_Z\sigma_{YZ} \\ {\beta^*_Y}^2\sigma^2_{\varepsilon_Y} &= {\beta^*_Y}^2\sigma^2_Y - \beta^*_Y\sigma_{XY} - \beta^*_Y\beta^*_Z\sigma_{YZ} + \beta^*_Z\sigma_{XZ} \\ {\beta^*_Z}^2\sigma^2_{\varepsilon_Z} &= {\beta^*_Z}^2\sigma^2_Z - \beta^*_Z\sigma_{XZ} - \beta^*_Z\beta^*_Y\sigma_{ZY} + \beta^*_Y\sigma_{XY} \end{aligned} \tag{2.A.14}$$

Expressing the rescaling coefficients β^*_Y and β^*_Z in (2.A.14) as covariance ratios (see (2.A.12)), and correcting for the rescaling coefficients on the left hand side of the equation further yields:

$$\begin{aligned} \sigma^2_{\varepsilon_X} &= \sigma^2_X - \frac{\sigma_{XY}\sigma_{XZ}}{\sigma_{YZ}} \\ \sigma^2_{\varepsilon_Y} &= \sigma^2_Y - \frac{\sigma_{YX}\sigma_{YZ}}{\sigma_{XZ}} \\ \sigma^2_{\varepsilon_Z} &= \sigma^2_Z - \frac{\sigma_{ZX}\sigma_{ZY}}{\sigma_{XY}} \end{aligned} \tag{2.A.15}$$

which is identical to the error variance estimates in (2.A.11), which were obtained using the covariance notation. Note that this identity only holds if the TC-based rescaling - as proposed by *Stoffelen* (1998) - is used as the a priori rescaling in the difference notation. However, the use of any other rescaling method can be considered as non-optimal TC implementation.

2.B The impact of representativeness errors on TC error variance estimates

Here we will demonstrate the impact of differing spatial representativeness of the data sets by means of differing underlying soil moisture signal components using the covariance notation. Note that this approach is not different to other approaches, where the correlated signal components are considered as cross-correlated random errors in the data sets (*Stoffelen*, 1998; *Vogelzang and Stoffelen*, 2012). Two different cases will be distinguished: (i) one point-scale in situ measurement together with two coarse-scale data sets that have a comparable spatial representativeness, and (ii) three data sets with significantly different spatial representativeness. The first scenario will be denoted as two-scale problem and the second as three-scale problem.

2.B.1 Two-scale problem

In this scenario we consider a soil moisture signal within an area A_j that is jointly observed by all three data sets, and an additional signal within an area A_c that is common only to the two coarse-scale data sets. The coarse-scale data sets are therefore assumed to sample the entire area $A_j + A_c$ while the point scale observations are assumed to be representative for A_j only. The signal within A_c originates from soil moisture variations that do not take place at the point location (e.g., localized rainfall events). Both signal components are assumed to be orthogonal and mutually uncorrelated. The point-scale data set will be denoted as X, and the coarse-scale data sets as Y and Z, respectively. Following (2.5), their variances can be written as:

$$\begin{aligned}
\sigma_X^2 &= \beta_X^2 \sigma_{\Theta_j}^2 + \sigma_{\varepsilon_X}^2 \\
\sigma_Y^2 &= \beta_Y^2 \frac{1}{(A_j + A_c)^2} (A_j^2 \sigma_{\Theta_j}^2 + A_c^2 \sigma_{\Theta_c}^2) + \sigma_{\varepsilon_Y}^2 \\
\sigma_Z^2 &= \beta_Z^2 \frac{1}{(A_j + A_c)^2} (A_j^2 \sigma_{\Theta_j}^2 + A_c^2 \sigma_{\Theta_c}^2) + \sigma_{\varepsilon_Z}^2
\end{aligned} \tag{2.B.1}$$

and their covariances as:

$$\begin{aligned}
\sigma_{XY} &= \beta_X \beta_Y \frac{A_j}{(A_j + A_c)} \sigma_{\Theta_j}^2 \\
\sigma_{XZ} &= \beta_X \beta_Z \frac{A_j}{(A_j + A_c)} \sigma_{\Theta_j}^2 \\
\sigma_{YZ} &= \beta_Y \beta_Z \frac{1}{(A_j + A_c)^2} (A_j^2 \sigma_{\Theta_j}^2 + A_c^2 \sigma_{\Theta_c}^2)
\end{aligned} \tag{2.B.2}$$

where $\sigma_{\Theta_j}^2$ is the joint soil moisture signal that is observed by all three data sets, and $\sigma_{\Theta_c}^2$ is the soil moisture signal that is observed only by the coarse-scale data sets. The error variance estimator

(2.6) therefore expands into:

$$\begin{aligned}
\sigma_X^2 - \frac{\sigma_{XY}\sigma_{XZ}}{\sigma_{YZ}} &= \sigma_{\varepsilon_X}^2 + \beta_X^2\sigma_{\Theta_j}^2\left(1 - \frac{1}{1+\frac{A_c^2\sigma_{\Theta_c}^2}{A_j^2\sigma_{\Theta_j}^2}}\right)\\
\sigma_Y^2 - \frac{\sigma_{YX}\sigma_{YZ}}{\sigma_{XZ}} &= \sigma_{\varepsilon_Y}^2\\
\sigma_Z^2 - \frac{\sigma_{ZX}\sigma_{ZY}}{\sigma_{XY}} &= \sigma_{\varepsilon_Z}^2
\end{aligned} \tag{2.B.3}$$

That is, TC will penalize the point-scale data set for its limited representativeness at the coarse scale, whereas no representativeness error is assigned to the error estimates of the coarse-scale data sets. Note that the representativeness error also depends on the fraction between A_c and A_j. The larger the area for which the point scale measurement is representative, the smaller is the impact of representativeness errors, even if the soil moisture variations that are observed only by the coarse-scale data sets are very strong.

2.B.2 Three-scale problem

In this scenario we have three data sets of significantly different spatial representativeness such as a point-scale in situ-measurement, a medium-scale land surface model, and a coarse-scale satellite data set. These will be - in the same order - denoted as X, Y, and Z, respectively. The coarse-scale observations are assumed to sample the entire area $A_j + A_m + A_c$, the medium-scale observations are assumed to sample only the area $A_j + A_m$ and the point observations are assumed to be representative for the area A_j only. The soil moisture signals within A_j, A_m, and A_c are again assumed to be orthogonal and mutually uncorrelated. The variances of the data sets are given as:

$$\begin{aligned}
\sigma_X^2 &= \beta_X^2\sigma_{\Theta_j}^2 + \sigma_{\varepsilon_X}^2\\
\sigma_Y^2 &= \beta_Y^2\frac{1}{(A_j+A_m)^2}(A_j^2\sigma_{\Theta_j}^2 + A_m^2\sigma_{\Theta_m}^2) + \sigma_{\varepsilon_Y}^2\\
\sigma_Z^2 &= \beta_Z^2\frac{1}{(A_j+A_m+A_c)^2}(A_j^2\sigma_{\Theta_j}^2 + A_m^2\sigma_{\Theta_m}^2 + A_c^2\sigma_{\Theta_c}^2) + \sigma_{\varepsilon_Z}^2
\end{aligned} \tag{2.B.4}$$

and their covariances as:

$$\begin{aligned}
\sigma_{XY} &= \beta_X\beta_Y\frac{A_j}{(A_j+A_m)}\sigma_{\Theta_j}^2\\
\sigma_{XZ} &= \beta_X\beta_Z\frac{A_j}{(A_j+A_m+A_c)}\sigma_{\Theta_j}^2\\
\sigma_{YZ} &= \beta_Y\beta_Z\frac{1}{(A_j+A_m)(A_j+A_m+A_c)}(A_j^2\sigma_{\Theta_j}^2 + A_m^2\sigma_{\Theta_m}^2)
\end{aligned} \tag{2.B.5}$$

where $\sigma^2_{\Theta_j}$ is again the joint soil moisture signal observed by all three data sets, $\sigma^2_{\Theta_m}$ is the soil moisture signal that is common to the medium- and the coarse-scale data set, and $\sigma^2_{\Theta_c}$ is the soil moisture signal that is observed by the coarse-scale data set only. All three signal components are assumed to be orthogonal. In this case, the error variance estimator (2.6) expands into:

$$
\begin{aligned}
\sigma^2_X - \frac{\sigma_{XY}\sigma_{XZ}}{\sigma_{YZ}} &= \sigma^2_{\varepsilon_X} + \beta^2_X \sigma^2_{\Theta_j} \left(1 - \frac{1}{1 + \frac{A^2_m \sigma^2_{\Theta_m}}{A^2_j \sigma^2_{\Theta_j}}} \right) \\
\sigma^2_Y - \frac{\sigma_{YX}\sigma_{YZ}}{\sigma_{XZ}} &= \sigma^2_{\varepsilon_Y} \\
\sigma^2_Z - \frac{\sigma_{ZX}\sigma_{ZY}}{\sigma_{XY}} &= \sigma^2_{\varepsilon_Z} + \beta^2_Z \frac{A^2_c}{(A_j + A_m + A_c)^2} \sigma^2_{\Theta_c}
\end{aligned}
\tag{2.B.6}
$$

While TC still penalizes the point-scale data set - this time for its limited representativeness at the medium scale - it additionally assigns a representativeness error to the coarse-scale data set, represented by its sensitivity to the soil moisture variations that are not observed by the other data sets. The error variance estimate for the medium-scale data set remains unbiased. Note that the representativeness errors again depend on the relative fraction of the areas for which the different data sets are representative.

Chapter 3

The potential of 2D Kalman filtering for soil moisture data assimilation

We examine the potential for parameterizing a two-dimensional (2D) land data assimilation system using spatial error auto-correlation statistics gleaned from a triple collocation analysis and the triplet of: (1) active microwave-, (2) passive microwave- and (3) land surface model-based surface soil moisture products. Results demonstrate that, while considerable spatial error auto-correlation exists in the errors for all three products, the inclusion of this information into a 2D assimilation system does not significantly improve the performance of the system relative to a one-dimensional (1D) baseline. This result is explained via an analytical evaluation of the impact of spatial error auto-correlation on the steady-state Kalman gain, which reveals that 2D filtering requires the existence of large auto-correlation differences (between the assimilation model and the assimilated observations) in order to enhance the analysis relative to a 1D filtering baseline. As a result, large error auto-correlations alone (in both the model or observations) are not sufficient to motivate the application of a 2D land assimilation system. These results have important consequences for the development of land data assimilation systems designed to ingest satellite derived surface soil moisture products for water resource and climate applications.

---

* This chapter is an edited version of: Gruber, A., Crow, W., Dorigo, W., & Wagner, W. (2015). The potential of 2D Kalman filtering for soil moisture data assimilation. Remote Sensing of Environment, 171, 137-148.

3.1 Introduction

Most current applications of satellite-based surface soil moisture retrievals involve their assimilation into a continuous surface water balance model to produce a merged (model/remote-sensing) soil moisture analysis product (*Bolten and Crow*, 2012; *Brocca et al.*, 2012; *de Rosnay et al.*, 2013). The accuracy of such analyses depend in part on the appropriate statistical parameterization of errors in both the diagnostic remote sensing retrievals and the prognostic water balance model (*Crow and Van Loon*, 2006; *Reichle et al.*, 2008). To date, most land data assimilation systems have been based on relatively crude and approximate treatment of such errors. This error parameterization challenge becomes even greater when data assimilation are expanded from a one-dimensional (1D) form (where observations are used to update only horizontally-collocated model states) to a two-dimensional (2D) structure (where observations of a particular grid cell are horizontally-transported to update the state in neighboring grid cells as well). In such a 2D approach, the spatial auto-correlation of observation and modeling errors must be parameterized in addition to their 1D variances (*Reichle and Koster*, 2003).

Recent progress has been made in leveraging the simultaneous acquisition of surface soil moisture retrievals acquired from active and passive microwave observations to provide an improved error parameterization for a 1D data assimilation system (*Crow and Yilmaz*, 2014). The basis of such approaches is the application of so-called triple collocation (TC) analysis to a triplet of soil moisture estimates acquired from three independent sources (typically active microwave, passive microwave, and soil moisture estimates derived from a water-balance model) (*Stoffelen*, 1998; *Scipal et al.*, 2008; *Dorigo et al.*, 2010). However, relatively little is known about the potential for applying TC to estimate spatial error auto-correlation parameters required to parameterize a 2D filter. Even more fundamentally, the role of spatial error auto-correlation in determining the relative advantage of 2D approaches (versus a simpler 1D baseline) has not fully been explored in land data assimilation.

This analysis will leverage the simultaneous availability of surface soil moisture retrievals from: the active-microwave-based Advanced SCATterometer (ASCAT) instrument (*Wagner et al.*, 1999; *Naeimi*, 2009) and the passive-microwave-based Soil Moisture and Ocean Salinity (SMOS) mission (*Kerr et al.*, 2010) to explore the potential advantages of parameterizing a 2D data assimilation system using error parameters gleaned from a TC analysis. Special emphasis will be placed on the ability of a TC-based system to acquire sufficiently accurate spatial error auto-correlation parameters to motivate the expansion of a (simpler) 1D data assimilation into a full 2D approach. It should be mentioned that the presence of non-zero temporal observation error auto-correlation is neglected in this study as its impact on data assimilation systems has already been investigated by a large number of studies (*Crow and Van den Berg*, 2010). However, the TC-based system presented here does not require temporally white errors in order to provide unbiased results, therefore our findings are directly transferable to the presumably more realistic case of error auto-correlation in both space and time.

The 2D Kalman filter and the TC-based method for estimating spatial error cross-correlations are described in Section 3.2. A synthetic and a real data experiment are employed in Section 3.3. Section 3.4 provides an analytical investigation of the obtained findings. A summary and conclusions are provided in Section 3.5.

3.2 Background

3.2.1 2D Kalman filter

Our methodology is based on the application of 1D and 2D data assimilation approaches to assimilate surface soil moisture observations into a simple auto-regressive soil moisture model:

$$\theta^m_{x,t} = \gamma_x \theta^m_{x,t-1} + P^m_{x,t} \tag{3.1}$$

where θ^m is the modeled soil moisture state; x is the spatial location; t is the daily time index; γ is a dimensionless loss variable, and P^m is a satellite-based estimate of accumulated daily rainfall. Note that - despite their extreme simplicity - simple auto-regressive models like (3.1) perform just as well as more complex (nonlinear) models when applied to large-scale applications such as agricultural drought monitoring (*Crow et al.*, 2012a). In addition, they provide a transparent basis for evaluating new land data assimilation methodologies. For convenience, all soil moisture and precipitation quantities referenced below are considered to be seasonal anomalies obtained by first subtracting out a seasonal climatological cycle. Note that, due to its linearity, this anomaly decomposition does not affect the validity of (3.1).

A single, spatially-distributed set of remotely-sensed soil moisture (anomaly) retrievals (θ^o) are then assimilated into (3.1). Both model and observed soil moisture anomalies are assumed to have a linear relationship with (unknown) true soil moisture anomalies (θ):

$$\begin{aligned} \theta^m_{x,t} &= \theta_{x,t} + \varepsilon^m_{x,t} \\ \theta^o_{x,t} &= \beta^o_x(\theta_{x,t} + \varepsilon^o_{x,t}) \end{aligned} \tag{3.2}$$

where β^o is a constant scaling factor to correct for systematic errors in the observations, and ε^m and ε^o are Gaussian-distributed, zero-mean errors in the model-estimates and observed surface soil moisture, respectively. Note that (3.2) follows the required data assimilation assumption that the model lacks systematic errors. Therefore, prior to data assimilation, observations $\theta^o_{x,t}$ must be rescaled by a factor of $(\beta^o_x)^{-1}$ to produce a bias-corrected set of anomaly observations $\hat{\theta}^o_{x,t}$ which are systematically consistent with the model .

The covariance matrices for model forecast noise (**Q**) and (rescaled) observation error (**R**) are then given as:

$$\mathbf{Q} = (1 - \gamma_x^2)\langle \vec{\varepsilon}^{\,m} \vec{\varepsilon}^{\,m\top} \rangle \qquad \mathbf{R} = \langle \hat{\vec{\varepsilon}}^{\,o} \hat{\vec{\varepsilon}}^{\,o\top} \rangle \tag{3.3}$$

where the brackets $\langle \cdot \rangle$ denote temporal averaging and the error vectors $\vec{\varepsilon} = (\varepsilon_{x_1} \cdots \varepsilon_{x_n})^\top$ contain the errors in model estimates and observations at n adjacent spatial locations.

Note that **Q** represents the variance of the white noise that is added during each forecast step of the model in (3.1) - due to the random error in the precipitation data set P^m - whereas the $\langle \vec{\varepsilon}^{\,m} \vec{\varepsilon}^{\,m\top} \rangle$ term alone represents the integrative steady-state impact of this noise on $\theta^m_{x,t}$, i.e., the red noise variance of the model due to its auto-regressive structure.

The spatial error covariance matrix of the model background forecast is given as:

$$\mathbf{M}_t = \vec{\gamma} \mathbf{M}_{t-1} \vec{\gamma}^\top + \mathbf{Q} \tag{3.4}$$

where $\vec{\gamma} = (\gamma_{x_i} \cdots \gamma_{x_n})^\top$. Using the (rescaled) observation state vector $\vec{Y}_t = (\hat{\theta}^o_{x_1,t} \cdots \hat{\theta}^o_{x_n,t})^\top$, the model state vector $\vec{X}_t = (\theta^m_{x_1,t} \cdots \theta^m_{x_n,t})^\top$, and - as only one observation data set is assimilated - an $n \times n$ identity matrix as the observation operator **H**, we can write the Kalman gain as:

$$\mathbf{K}_t = \mathbf{M}_t \mathbf{H}^\top (\mathbf{H} \mathbf{M}_t \mathbf{H}^\top + \mathbf{R})^{-1} \tag{3.5}$$

and the state and forecast error covariance update equations as:

$$\begin{aligned} \vec{X}_t' &= \vec{X}_t + \mathbf{K}_t (\vec{Y}_t - \mathbf{H} \vec{X}_t) \\ \mathbf{M}_t' &= (\mathbf{I} - \mathbf{K}_t \mathbf{H}) \mathbf{M}_t \end{aligned} \tag{3.6}$$

These equations are developed here for the simple case of assimilating only one type of soil moisture observations; however, they are readily scalable to the case of multiple observation data sets.

3.2.2 Parameter estimation

For the 1D Kalman filter with k observations, one has to estimate $2k+1$ parameters prior to filtering - namely the scaling coefficients between the model and the observations - and the model and observation error variances (i.e., $(\beta^o_x)^{-1}$, Q_{xx}, and R_{xx}). The 2D case further requires the knowledge of the spatial error auto-covariances, which entails the estimation of one additional parameter for each assimilated neighboring pixel for each data set, respectively.

Triple collocation (TC) (*Stoffelen*, 1998) has been previously applied in 1D Kalman filtering to provide unbiased estimates of model and observation error variances and relative scaling coefficients (*Crow and Yilmaz*, 2014). It requires exactly three collocated data sets with orthogonal errors and zero error cross-correlation. These requirements are commonly assumed to be met when using a soil moisture triplet composed of: 1) active microwave satellite retrievals (θ^a), 2) passive retrievals (θ^p) and 3) model-based soil moisture estimates (θ^m). However, to date, TC has not been applied to estimate spatial error auto-covariance statistics. Therefore, our goal here is the application of TC to estimate the entire observation error covariance matrices ($\mathbf{R}^a$ and $\mathbf{R}^p$) and model forecast noise matrix ($\mathbf{Q}$) required to parameterize the 2D Kalman filter described in Section 3.2.1. Note that this requires the estimation of both diagonal (i.e., variance) and off-diagonal (i.e., spatial auto-covariance) matrix components.

Here we will apply the covariance notation of *Stoffelen* (1998) to the TC problem. By using the error model in (3.2), the respective model and observation variances and covariances at a single coinciding spatial location can be written as:

$$\begin{aligned}\mathrm{Var}(\theta^i_x) &= {\beta^i_x}^2\mathrm{Var}(\theta_x) + {\beta^i_x}^2\mathrm{Var}(\varepsilon^i_x)\\ \mathrm{Cov}(\theta^i_x,\theta^j_x,) &= \beta^i_x\beta^j_x\mathrm{Var}(\theta_x)\end{aligned} \tag{3.7}$$

where i and j denote either the model (m) or an active (a) or passive (p) observation data set. Combining the covariances of all three data sets (i, j, and k) allows us to estimate the (individually biased) true signal variance (*McColl et al.*, 2014):

$${\beta^i_x}^2\mathrm{Var}(\theta_x) = \frac{\mathrm{Cov}(\theta^i_x,\theta^j_x)\mathrm{Cov}(\theta^i_x,\theta^k_x)}{\mathrm{Cov}(\theta^j_x,\theta^k_x)} \tag{3.8}$$

Following the typical data assimilation assumption that the model lacks systematic error (i.e., $\beta^m = 1$), estimates of the scaling coefficients - required to correct for the systematic error in the observations - can be obtained as:

$$\begin{aligned}(\beta^a_x)^{-1} &= \frac{\mathrm{Cov}(\theta^m_x,\theta^p_x)}{\mathrm{Cov}(\theta^a_x,\theta^p_x)}\\ (\beta^p_x)^{-1} &= \frac{\mathrm{Cov}(\theta^m_x,\theta^a_x)}{\mathrm{Cov}(\theta^p_x,\theta^a_x)}\end{aligned} \tag{3.9}$$

Subtracting the (biased) true signal variance - as obtained from (3.8) - from the respective total data set variance in (3.7) and correcting for the bias using (3.9) yields unbiased estimates of the individual error variances as:

$$
\begin{aligned}
Q_{xx} &= (1-\gamma_x^2)\mathrm{Var}(\varepsilon_x^m) = (1-\gamma_x^2)\left[\mathrm{Var}(\theta_x^m) - \frac{\mathrm{Cov}(\theta_x^m,\theta_x^a)\mathrm{Cov}(\theta_x^m,\theta_x^p)}{\mathrm{Cov}(\theta_x^a,\theta_x^p)}\right] \\
R_{xx}^a &= \mathrm{Var}(\varepsilon_x^a) = (\beta_x^a)^{-2}\left[\mathrm{Var}(\theta_x^a) - \frac{\mathrm{Cov}(\theta_x^a,\theta_x^m)\mathrm{Cov}(\theta_x^a,\theta_x^p)}{\mathrm{Cov}(\theta_x^m,\theta_x^p)}\right] \\
R_{xx}^p &= \mathrm{Var}(\varepsilon_x^p) = (\beta_x^p)^{-2}\left[\mathrm{Var}(\theta_x^p) - \frac{\mathrm{Cov}(\theta_x^p,\theta_x^m)\mathrm{Cov}(\theta_x^p,\theta_x^a)}{\mathrm{Cov}(\theta_x^m,\theta_x^a)}\right]
\end{aligned}
\tag{3.10}
$$

Note that even though we assume that the model lacks systematic error (i.e., $\beta^m = 1$), we need to correct the model error variance estimate for the model-memory through $(1-\gamma_x^2)$.

In a similar way, we can obtain the spatial auto-covariances between errors at two adjacent locations x_1 and x_2 from the spatial covariances of a single data set i and between two data sets i and j, respectively, where i and j correspond again to either active (a), passive (p), or model (m) derived products:

$$
\begin{aligned}
\mathrm{Cov}(\theta_{x_1}^i,\theta_{x_2}^i) &= \beta_{x_1}^i\beta_{x_2}^i\mathrm{Cov}(\theta_{x_1},\theta_{x_2}) + \beta_{x_1}^i\beta_{x_2}^i\mathrm{Cov}(\varepsilon_{x_1}^i,\varepsilon_{x_2}^i) \\
\mathrm{Cov}(\theta_{x_1}^i,\theta_{x_2}^j) &= \beta_{x_1}^i\beta_{x_2}^j\mathrm{Cov}(\theta_{x_1},\theta_{x_2})
\end{aligned}
\tag{3.11}
$$

Just like the error variances before, the spatial error auto-covariances of the model and the observations can be obtained from (3.11) as:

$$
\begin{aligned}
Q_{x_1x_2} &= \sqrt{(1-\gamma_{x_1}^2)(1-\gamma_{x_2}^2)}\left[\mathrm{Cov}(\theta_{x_1}^m,\theta_{x_2}^m) - \frac{\mathrm{Cov}(\theta_{x_1}^m,\theta_{x_2}^a)\mathrm{Cov}(\theta_{x_1}^p,\theta_{x_2}^m)}{\mathrm{Cov}(\theta_{x_1}^p,\theta_{x_2}^a)}\right] \\
R_{x_1x_2}^a &= (\beta_{x_1}^a\beta_{x_2}^a)^{-1}\left[\mathrm{Cov}(\theta_{x_1}^a,\theta_{x_2}^a) - \frac{\mathrm{Cov}(\theta_{x_1}^a,\theta_{x_2}^m)\mathrm{Cov}(\theta_{x_1}^p,\theta_{x_2}^a)}{\mathrm{Cov}(\theta_{x_1}^p,\theta_{x_2}^m)}\right] \\
R_{x_1x_2}^p &= (\beta_{x_1}^p\beta_{x_2}^p)^{-1}\left[\mathrm{Cov}(\theta_{x_1}^p,\theta_{x_2}^p) - \frac{\mathrm{Cov}(\theta_{x_1}^p,\theta_{x_2}^m)\mathrm{Cov}(\theta_{x_1}^a,\theta_{x_2}^p)}{\mathrm{Cov}(\theta_{x_1}^a,\theta_{x_2}^m)}\right]
\end{aligned}
\tag{3.12}
$$

In this way, TC can be applied to estimate all of the statistical parameters required to parameterize a 2D Kalman filter.

Note that (3.12) is identical to the method of triple collocation based temporal error auto-covariance estimation proposed by *Zwieback et al.* (2012) - except that the temporal lag has been replaced by the spatial lag. Spatial error auto-correlations of the model (C_Q) and the active (C_{R^a}) and passive (C_{R^p}) observations can be further derived as $C_Q = Q_{x_1x_2}(Q_{x_1x_1}Q_{x_2x_2})^{-\frac{1}{2}}$, $C_{R^a} = R_{x_1x_2}^a(R_{x_1x_1}^a R_{x_2x_2}^a)^{-\frac{1}{2}}$, and $C_{R^p} = R_{x_1x_2}^p(R_{x_1x_1}^p R_{x_2x_2}^p)^{-\frac{1}{2}}$. The method will be validated in Section 3.3.1.1.

3.3 Demonstration

In this section we will evaluate the parameterization of a 2D Kalman filter using the TC-based error parameter estimates introduced in Section 3.2.2. Evaluations will be based on both synthetic identical twin experiments (Section 3.3.1) and a real data analysis evaluated using comparisons against independent ground-based surface soil moisture measurements (Section 3.3.2). Particular attention will be paid to the added skill associated with applying a 2D filtering strategy versus a 1D baseline approach which ignores the presence of spatial auto-correlation in modeling and observation errors. Only one type of satellite-based observation will be assimilated here since the simultaneous assimilation of two observation data sets gives rise to additional issues (e.g., the accurate characterization of error cross-correlation) which are beyond the scope of this paper. However, the second observation data set is needed as an instrument in the TC analysis in order to estimate the error variance and auto-covariance structures of both the model and the observation data set to be assimilated.

3.3.1 Synthetic experiment

Synthetic experiments were based on: 1) the generation of a soil moisture reference product representing the truth via an unperturbed integration of (3.1), 2) the artificial perturbation of this product using noise with covariance $\mathbf{R}$ to generate synthetic observations, and 3) the re-assimilation of these synthetic observations back into (3.1) after the model has been artificially perturbed using forecast noise consistent with $\mathbf{Q}$. True soil moisture products generated in the first step contained values for one center pixel and neighboring pixels in each of the four cardinal directions. In this way, the 2D assimilation problem is effectively localized within a single pixel length in each cardinal direction.

Results were generated for a large number of different cases. In particular, the spatial error auto-correlation levels of the model and the observations (C_Q and C_R) were systematically varied between 0.05 [-] and 0.95 [-] in increments of 0.05 [-]. In addition, three separate cases of different relative error levels between the center and the neighboring pixels of the model and the observations (expressed as error ratios $\delta Q = Q_{x_2x_2} \cdot (Q_{x_1x_1})^{-1}$ and $\delta R = R_{x_2x_2} \cdot (R_{x_1x_1})^{-1}$, where x_1 denotes the center pixel and x_2 denotes the neighboring pixel) are investigated, namely $\delta Q = 0.8/1.0/1.2$ and $\delta R = 1.2/1.0/0.8$. The sample size of the data sets was 5000 days. The absolute error variances of the model and the observation data sets were held fixed at 110 mm^2 and 450 mm^2, respectively, so that the correlation of the synthetic data sets with respect to the reference (i.e., the synthetic truth) were comparable to those observed at the watershed sites in the real data experiment (see Section 3.3.2). Combined, these synthetic cases required the generation of 16245 separate synthetic data sets.

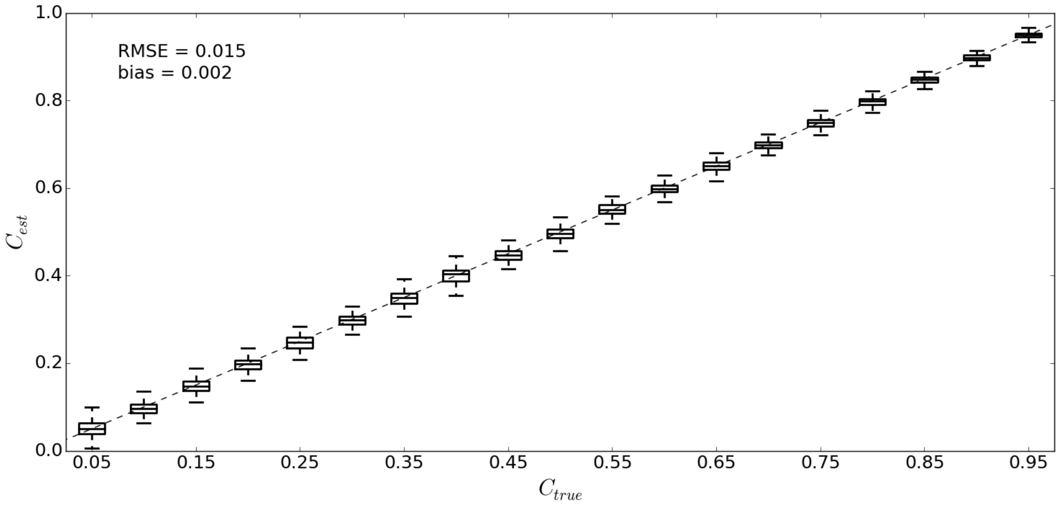

Figure 3.1: *The median and the inter-quartile-range (IQR) of estimated spatial error auto-correlations (y-axis) for different true auto-correlation levels (x-axis). The whisker length is 1.5*IQR. The dashed line is the 1:1 line.*

3.3.1.1 Estimation of spatial error auto-correlation

As an initial verification, Figure 3.1 shows the estimation accuracy of the TC based error auto-correlation estimates for different model and observation error auto-correlation levels. C_{est} represents the C_Q, C_{R^a}, and C_{R^p} estimates - as derived from (3.10) and (3.12) - of all 16245 synthetic data sets. The corresponding synthetically generated true error auto-correlation levels (C_{true}) can be recovered with a negligible bias and a Root-Mean-Square-Error (RMSE) below 0.02 [-]. Therefore, the application of TC to accurately estimate error auto-correlation information appears plausible. The precision of the estimates is slightly higher at high error auto-correlation levels because of their non-linear dependency on the uncertainties of the error variance estimates, which originates from the conversion of error auto-covariances to error auto-correlations. The precision of error auto-covariance estimates alone does not show such dependency on absolute error auto-covariance level. Notice that the RMSE of the estimator is a function of sample size, which is in our case very large ($n = 5000$ days). Moreover, the results presented here are based on well-controlled synthetic observations which represent a "perfect" soil moisture signal perturbed with a Gaussian noise component. Hence, higher estimation uncertainties are expected when using (imperfect) real data.

3.3.1.2 Impact on soil moisture analysis accuracy

Here we evaluate the improvement in forecast skill with respect to the 1D filter implementation in terms of the correlation to the (synthetically generated) truth. The average correlation of the open

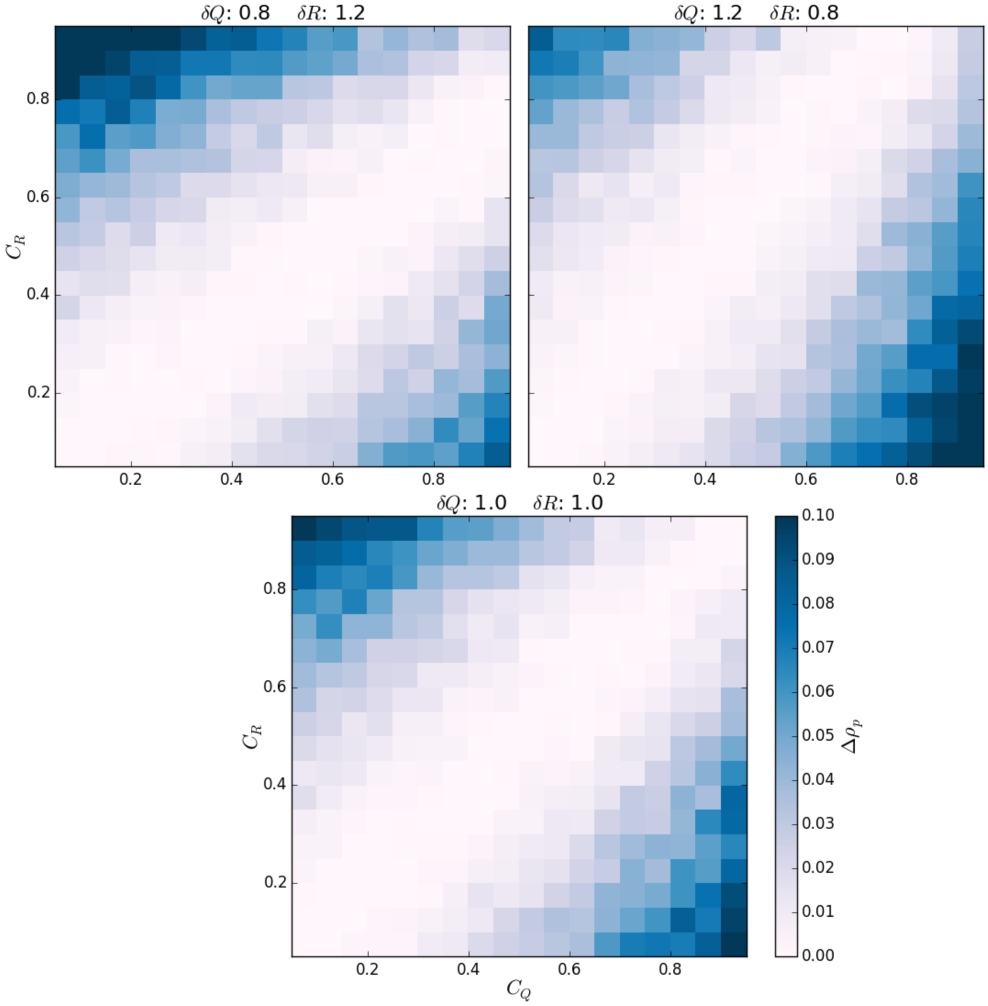

Figure 3.2: *Improvement of Pearson correlation (versus synthetic truth) of a 2D Kalman filtering analysis with respect to the 1D filtering case ($\Delta\rho_p$) for different δQ and δR. C_Q levels are plotted on the x-axis and C_R levels on the y-axis.*

loop run with the true time series is 0.53 [-], the average correlation between the truth and the 1D filter analysis is 0.75 [-]. Figure 3.2 shows the correlation improvement (above the 1D baseline) gained from 2D filtering as a function of C_Q, C_R,δQ, and δR.

The observed pattern in Figure 3.2 is somewhat counter-intuitive. From filter theory, one might expect that a high spatial error auto-correlation in either the model or the observation data set would lead to a higher weighting of the neighboring pixels of that particular data set, leading to an improvement with respect to a 1D filter. However, a significant skill improvement is only observed in cases where the model and the observation error auto-correlations have large relative differences (i.e., $|\Delta C| \gg 0$, where $\Delta C = C_R - C_Q$). Conversely, if $\Delta C \approx 0$ there is no skill improvement with respect to the 1D case, even if C_Q and C_R individually are close to unity.

This behavior further shows a certain dependency on differences between the error ratios of center and neighboring pixels of the model and the observations, respectively (i.e., differences between δQ and δR). If $\delta Q = \delta R$, then the skill improvement due to $\Delta C \neq 0$ is symmetrical around the $\Delta C = 0$ line with zero skill improvement on this line. Again contradicting intuition, an increase in the error level of a neighboring pixel of either of the data sets causes a larger improvement of the 2D filter skill with increasing spatial error auto-correlation of that data set, but only if the error level in the neighboring pixel of the other data set does not increase by the same factor.

In summary, the skill improvement of a 2D filter with respect to a 1D filter appears driven mainly by differences between the spatial error auto-correlations of the model versus the observations (i.e., ΔC) and by differences between the error ratios of center and neighboring pixels of the model and the observations, respectively (i.e., differences between δQ and δR), and not by the absolute values of these quantities. However, even if the additional assimilation of spatially displaced pixels was shown to potentially improve forecast skill, the maximum achievable improvement is only marginal. For the well-controlled synthetic cases presented here, the mean and maximum obtained correlation improvements (versus synthetic truth) over the entire ΔC range are 0.03 [-] and 0.16 [-], respectively.

Note that all synthetic results presented above are based on correctly-estimated error structures and statistics. Assigning large error auto-correlation differences where there are none due to estimation uncertainties when using real data (i.e., over- or underestimating the error auto-correlation in one data set) would further degrade 2D results and might even degrade 2D filter performance compared to the 1D case. Therefore, in order to provide a more thorough and realistic assessment, the next section examines the behavior of the 2D filter in a real data scenario. In addition, theoretical background for these results will be presented in Section 3.4.

3.3.2 Real data experiment

In addition to the synthetic experiment presented above, the data assimilation system described in Section 3.2 was also applied using actual satellite data. This section illustrates results for TC-based error auto-correlation estimates over the Contiguous United States (CONUS) and evaluates 2D filtering results (based on these estimates) over both four heavily-instrumented in situ watersheds which are highly representative at the coarse model scale and 228 single-site ground stations which are homogeneously distributed over CONUS.

The soil moisture model (θ^m) is driven by daily satellite-based rainfall estimates derived from Version 7 of the TMPA 3B42 algorithm (*Huffman et al.*, 2010). This algorithm generates TRMM-adjusted merged-infrared (IR) daily accumulated rainfall estimates using rainfall measuring instruments onboard the TRMM satellite - the Precipitation Radar (PR), the nine-channel passive microwave radiometer (TMI), and the five-channel Visible and Infrared Scanner (VIRS) - to adjust merged IR data which consists of GMS, GOES-E, GOES-W, Meteosat-7, Meteosat-5, and NOAA-12 data. Accumulated rainfall estimates are provided in mm/day at 00:00 UTC with a 0.25-degree by 0.25-degree spatial resolution.

The active satellite-based soil moisture estimates (θ^a) used here are those from the H-SAF H25 SM-OBS-4 Metop-A ASCAT time series product, retrieved using the TU Wien algorithm version WARP5.5 R2.1. Soil moisture is estimated as degree of saturation at a spatial resolution of 25 km, resampled to a 0.25-degree by 0.25-degree grid using a Hamming window approach. Measurements are provided with a temporal resolution of 1-3 days, according to the repeat cycle of ASCAT. The ASCAT Surface State Flag (SSF) (*Naeimi et al.*, 2012), derived from backscatter data using an empirical threshold-analysis algorithm, is used to detect and mask measurements taken under frozen or freezing/thawing conditions.

The passive satellite-based soil moisture estimates (θ^p) are extracted from the SMOS L3 product version 5.01, which provides global daily soil moisture maps resampled to 00:00 UTC and a spatial resolution of 0.25-degree by 0.25-degree. The original revisit time is 1-3 days, the temporal resolution at the center field of view (FOV) 35 km.

As already mentioned, only ASCAT retrievals are assimilated. Since such active microwave-based retrievals are assumed to be relatively unaffected by the orbit direction (i.e., the observation time), no distinction was made between retrievals from ascending and descending satellite orbits in order to increase the temporal measurement density. SMOS retrievals are solely used as the third data set in the TC analysis. As in the synthetic study, the 2D Kalman filter analysis (for a given pixel) is localized to consider only observations within that pixel or within any of the four neighboring pixels in each cardinal direction. Missing or invalid retrievals are treated as retrievals with very large errors so that their weights in the Kalman gain matrix vanish. That is, filtering updates were

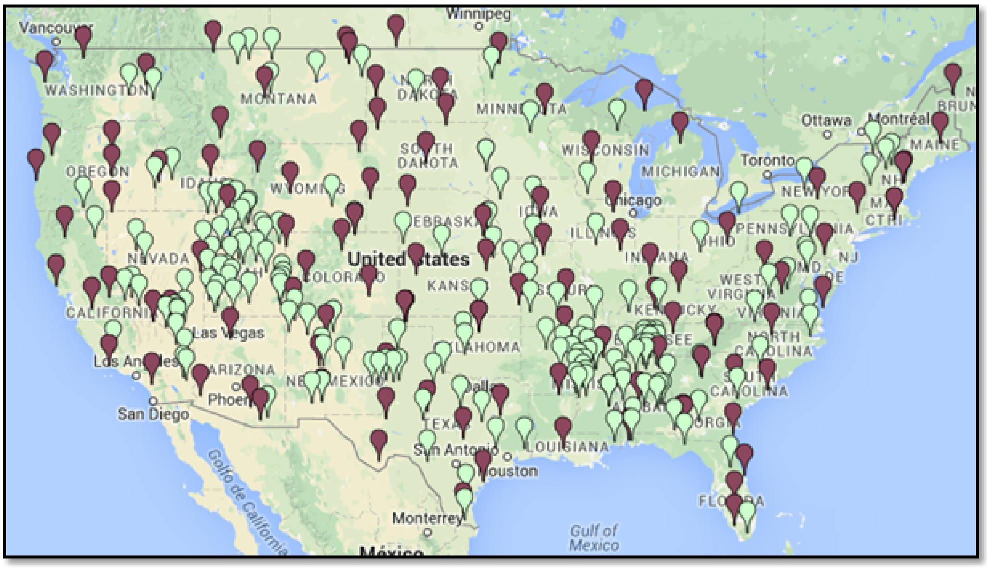

Figure 3.3: Station locations for the SCAN (green) and USCRN (red) soil moisture networks.

always performed if at least one valid pixel was available, irrespective of them being a center or a neighboring pixel.

The four watershed sites used were: the Reynolds Creek (RC) in Idaho, the Little Washita (LW) in Oklahoma, the Walnut Gulch (WG) in Arizona, and the Little River (LR) in Georgia. All four are operated as experimental watersheds by the United States Department of Agriculture's Agricultural Research Service (USDA ARS). Within each, multiple spatially-distributed ground-based surface (0-5 cm) soil moisture measurement were aggregated within the 0.25-degree pixels to provide a high-quality soil moisture representation at the satellite scale (*Jackson et al.*, 2010). The watersheds RC, LW, and WG provide representations for three, and LR for seven different 0.25-degree pixels. Their locations are indicated in subsequent figures.

The 228 single ground stations were drawn from the International Soil Moisture Network (ISMN; *Dorigo et al.*, 2011b) and are based on sites operated by the USDA Soil Climate Analysis Network (SCAN) and the U.S. Climate Reference Network (USCRN). Station locations are shown in Figure 3.3. Only surface soil moisture sensors that are placed within the first 10 cm of the soil were used. Single measurements have been masked based on an automated quality control procedure (*Dorigo et al.*, 2013).

3.3.2.1 Estimation of spatial error auto-correlation

Figure 3.4 shows the spatial error auto-correlation estimates of the TRMM-driven open-loop model run (C_Q) and ASCAT (C_R), as well as the error auto-correlation difference ($\Delta C = C_R - C_Q$) over CONUS for a lag distance of one pixel (i.e., 0.25-degrees). Individual correlations are calculated

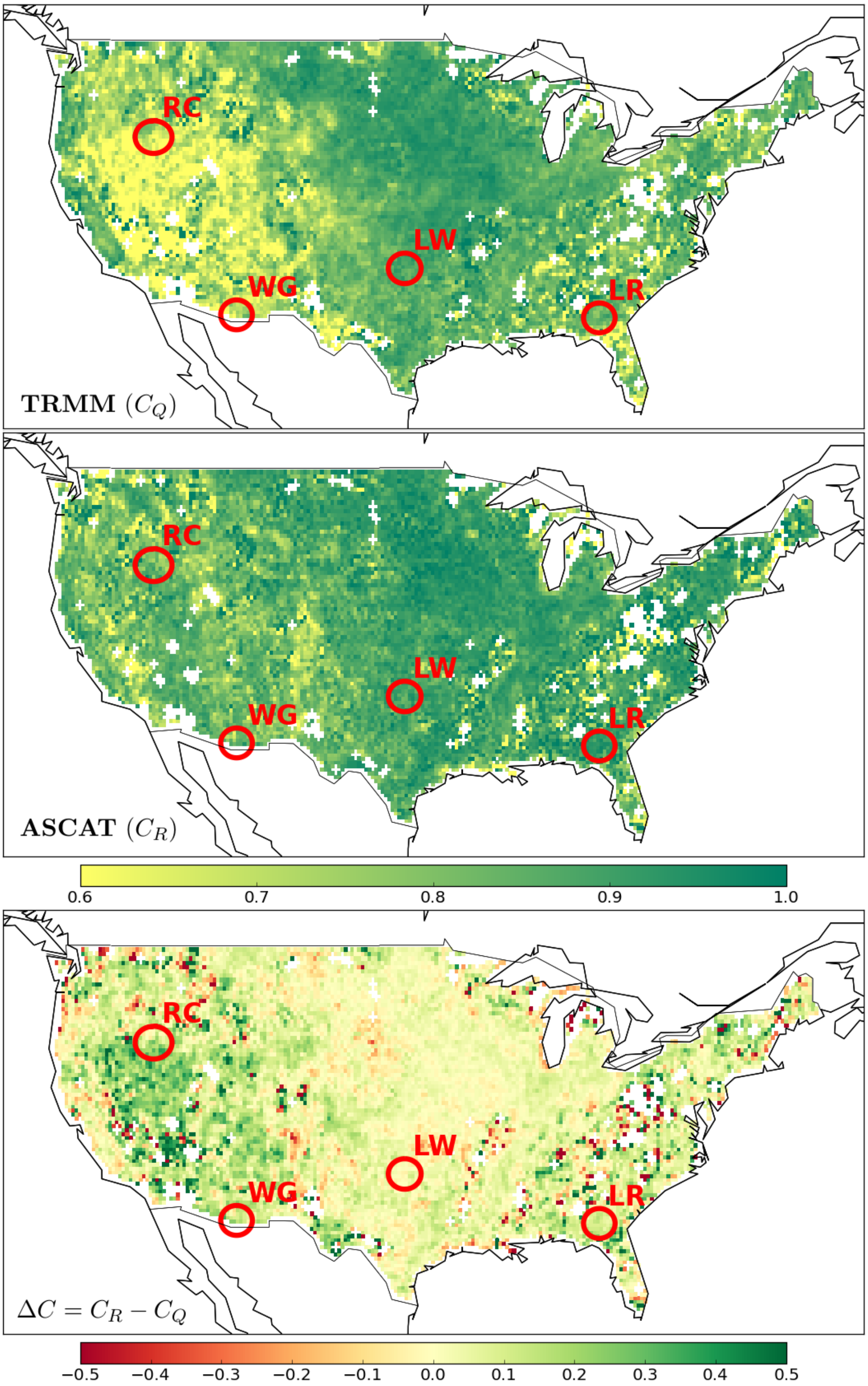

Figure 3.4: *Spatial error auto-correlations (averaged over all cardinal directions) for soil moisture estimates derived from TRMM using (3.1) (top) and for ASCAT surface soil moisture retrievals (middle), and the error auto-correlation difference between them (bottom). Red circles mark the USDA ARS watershed locations.*

in each of the four cardinal directions; however, for display purposes, values plotted in Figure 3.4 represent an omni-directional average. Overall, error auto-correlation levels of both soil moisture data sets are very high (typically > 0.8 [-]). The lower values of C_Q in the Western US are potentially related to increased topographic complexity, which leads to a much higher spatial heterogeneity in local precipitation patterns. High - and rather homogeneous - values of (C_R) can be explained by the spatial Hamming window resampling, which is applied in the retrieval algorithm. As a result, values of ΔC are typically close to zero with significantly non-zero areas concentrated mainly in the Western US. Figure 3.5 further shows estimates of the error variance ratios of the TRMM-based open-loop model run ($\delta Q = Q_{x_2x_2} \cdot (Q_{x_1x_1})^{-1}$), and of ASCAT ($\delta R = R_{x_2x_2} \cdot (R_{x_1x_1})^{-1}$), respectively. Their lower quartiles are 0.9 [-] and 0.9 [-], the medians are 1.0 [-] and 1.1 [-], and the upper quartiles are 1.2 [-] and 1.6 [-], respectively. The overall patterns follow those of the spatial error auto-correlations but appear to be noisier, which indicates generally higher sampling uncertainties in the TC based error variance estimates.

Taken at face value, the high spatial error auto-correlations in Figure 3.4 would seem to motivate the application of a 2D filter. However, synthetic experiment results in Section 3.3.1 suggest that the assimilation of neighboring pixels only leads to an improvement if the difference in the model and the observation error auto-correlation (i.e., ΔC) is large. The impact of these sampled error auto-correlations on real data assimilation results is examined numerically in Section 3.3.2.2 and analytically in Section 3.4.

3.3.2.2 Impact on soil moisture analysis accuracy

Figure 3.6 shows the correlation coefficients between the watershed-averages and the filtered satellite time series for the open-loop run lacking ASCAT assimilation and the 1D and 2D filtering cases after assimilating ASCAT. Despite the presence of significant spatial auto-correlation in TRMM and ASCAT errors (Figure 3.4), 2D filtering is not associated with a significant improvement in performance at any watershed site. At a few stations (RC-2, RC-3, WG-3, and LR-6), a slight (but statistically insignificant) improvement is apparent. However, in the majority of cases the skill of the 2D filter is almost the same as that of the 1D filter or, in some cases, even slightly lower (e.g., RC-1, WG-3, LR-2, LR-3, and LR-5). No relationship between skill improvement and the weight that is given to neighboring pixels is observed. Note that in theory, even if the neighboring pixels contain no information to update the state of the center pixel, their consideration in a 2D filtering analysis should not degrade the quality of the analysis with respect to the 1D filter. Instead such neighboring pixels should simply be assigned a weight of zero.

In order to evaluate results over a wider range of land surface conditions and measurement sites, Figure 3.7 plots correlation coefficients with all sites within the SCAN and USCRN networks

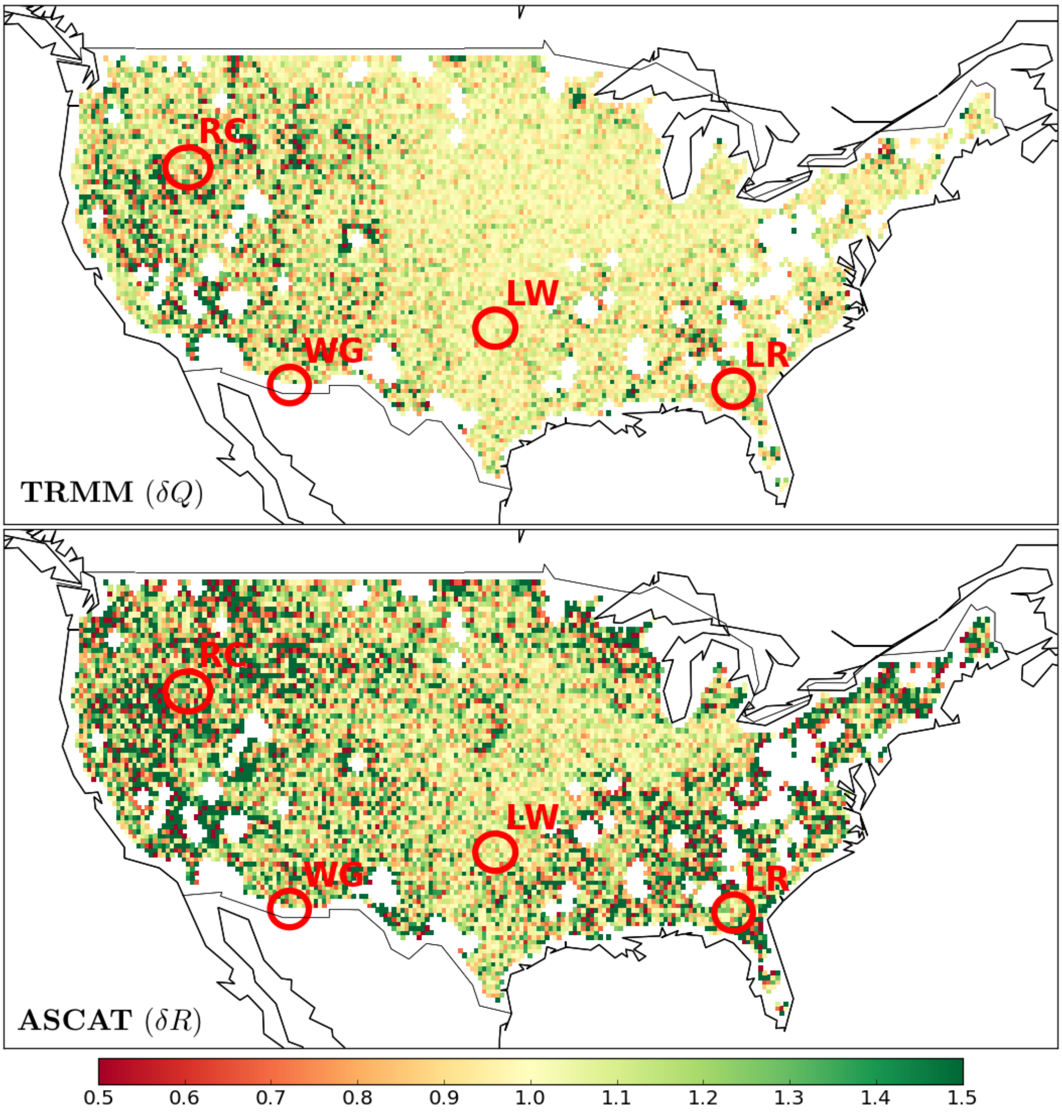

Figure 3.5: *Spatial error ratios (averaged over all cardinal directions) for soil moisture estimates derived from TRMM using (3.1) (top) and for ASCAT surface soil moisture retrievals (bottom). Red circles mark the USDA ARS watershed locations.*

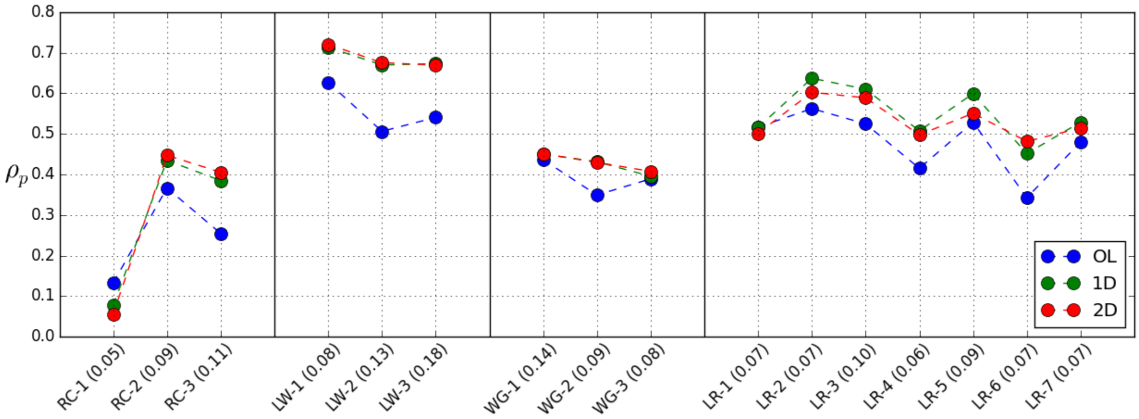

Figure 3.6: *Pearson correlation (ρ_p) between the open-loop model run without assimilating ASCAT observation (OL), the 1D Kalman filtering (1D) and 2D Kalman filtering (2D) estimates when assimilating ASCAT, and ground-based surface soil moisture observation results for individual 0.25-degree pixels located within each USDA ARS watershed site. Omnidirectionally averaged Kalman gain weights applied to observations within neighboring pixels are shown in brackets.*

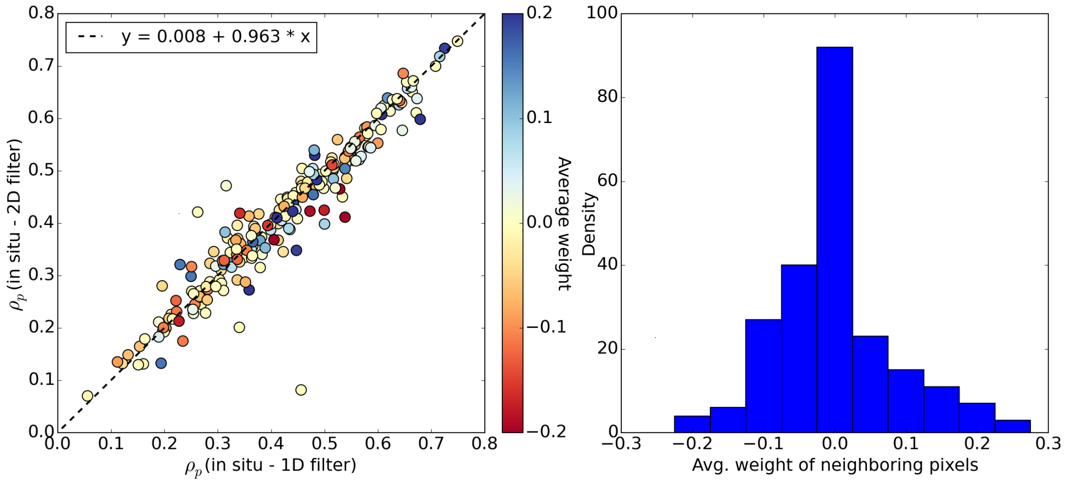

Figure 3.7: *Pearson correlation (ρ_p) between the soil moisture analysis and measurements acquired at SCAN and USCRN sites (left) for both the 1D filter case (x-axis) and the 2D filter case (y-axis), and a histogram of the average weights that are given to the neighboring pixels in the 2D filter case at the site locations (right).*

for both the 1D and 2D filter cases. Note that even if these sites might have a reduced coarse-scale representativeness compared to the watershed-averages, the relative performance increase or decrease of the different filtering approaches (i.e., 1D vs. 2D) should remain reliable in a statistical sense (*Liu et al.*, 2011a). The weights that are given to neighboring pixels, i.e., the first row of the steady-state expression of the Kalman gain in (3.5) for $t \gg 0$ (at each station averaged over all cardinal directions), are distributed from about -0.2 [-] to about +0.25 [-] with the majority of weights being close to zero, but with a slightly higher tendency for negative values. The average correlation of the 2D filter w.r.t. the ground stations remains with an ubRMSD of 0.05 [-] essentially unchanged relative to the 1D filter. As in the watershed case presented above, the magnitude of the weights does not correlate with the performance increase or decrease ($p < 0.01$).

In summary, despite the presence of large spatial auto-correlation in both modeling and observations errors little or no improvement (and in some cases even degradation) is noted upon the transition between 1D and 2D filtering. The lack of significant improvement (between 1D vs. 2D filtering) is related to the rather low weighting of neighboring pixels originating from very small error auto-correlation *differences*, and the noisy appearance of performance increase/decrease is likely related to the sensitivity of the weights to sampling errors in the error variance and error auto-covariance estimates. The following section explores these issues in greater analytical detail.

3.4 Analytical investigation of the Kalman gain behavior

Taken as a whole, the paradoxical finding of of Section 3.3 is that - despite the apparent presence of very high levels of error auto-correlation in (3.1) and in remotely-sensed surface soil moisture retrievals (Figure 3.4) - the assimilation of these retrievals into the model does not benefit significantly from the introduction of a 2D filtering strategy. Numerical results in Section 3.3.2 suggest that this lack of sensitivity results from (i) a lack of relative difference between the auto-correlation of modeling versus observation errors, and/or (ii) inaccuracies in estimated weights of neighboring pixels originating from uncertainties in TC-based error (co-)variance estimates.

In this section, we will develop an analytical explanation for this result. In particular, we will analytically investigate the expected impact of assimilating neighboring pixels by means of their given weight. For simplicity and without lacking generality we will restrict this consideration to one observation at the time $t = 0$, where the neighboring pixel at location y is used to update the pixel at location x. The optimal weighting that should be applied to the observation in pixel y (when updating the state in pixel x) is thus the off-diagonal element of the Kalman gain in (3.5) when replacing $\mathbf{M}^t$ with $\mathbf{Q}$ (i.e., the initial value in (3.4) at $t = 0$), which can be written as:

$$w_y = \frac{Q_{xy}(Q_{xx} + R_{xx}) - Q_{xx}(Q_{xy} + R_{xy})}{(Q_{xx} + R_{xx})(Q_{yy} + R_{yy}) - (Q_{xy} + R_{xy})^2} \tag{3.1}$$

Further justification for this simplification of the steady-state Kalman gain expression is provided in Appendix 3.A.

Let us now express the error variances at the two different locations x and y in terms of their ratio, i.e., $Q_{xx} \equiv Q$ and $Q_{yy} = Q\delta Q$ with $\delta Q = Q_{yy} \cdot (Q_{xx})^{-1}$ and likewise $R_{xx} \equiv R$ and $R_{yy} = R\delta R$ with $\delta R = R_{yy} \cdot (R_{xx})^{-1}$. Let us further express the spatial error auto-covariance of the model and the observation data set (i.e., the Q_{xy} and R_{xy}) in terms of a linear correlation coefficient (i.e., as C_Q and C_R) and these spatial error auto-correlations in terms of their relative difference, i.e., $Q_{xy} = C_Q Q\sqrt{\delta Q}$ and $R_{xy} = (C_Q + \Delta C)R\sqrt{\delta R}$ with $\Delta C = C_R - C_Q$. We can then rewrite (3.1) to:

$$w_y = \frac{QRC_Q(\sqrt{\delta Q} - \sqrt{\delta R} - QR\Delta C\sqrt{\delta R})}{(Q+R)(Q\delta Q + R\delta R) - (C_Q Q\sqrt{\delta Q} + C_Q R\sqrt{\delta R} + \Delta C R\sqrt{\delta R})^2} \tag{3.2}$$

In (3.2) the significant dependency of the 2D Kalman gain weights on differences in the error characteristics of the model and the observations - observed earlier in numerical results - becomes apparent. That is, if the error properties are very similar (i.e., $\delta Q \approx \delta R$ and $\Delta C \approx 0$), then the numerator in (3.2) will approach zero whereas the denominator will approach $(Q+R)^2(1 - {C_Q}^2)$. This behavior is illustrated in Figure 3.1. Even if δQ and δR slightly deviate from each other, the approximate $(1 - {C_Q}^2)$ proportionality in the denominator will cause w_y to remain almost zero for absolute C_Q (and C_R) values below 0.7 - 0.8 [-] (Figure 3.1a). For higher C_Q (and C_R) values, w_x becomes very sensitive to fluctuations in absolute error variance and covariance estimates (Figure 3.1b) but remains zero if $\Delta C = 0$ and $\delta Q = \delta R$. As a consequence, Kalman gain weights that are significantly different from zero and relatively robust against fluctuation in error parameter estimates can only be obtained if $\Delta C \gg 0$ (Figure 3.1c,d). Conversely, this implies that caution is required when implementing a 2D filter in all cases when ΔC is relatively small, and - given the low ΔC noted in real data results (Figure 3.4) - explains our lack of success in applying a 2D filter to real data cases (Figures 3.6 and 3.7).

The calculated simplified Kalman gain weights - as obtained from (3.1) - are shown in Figure 3.2. The patterns roughly follow the distribution of the error auto-correlation differences (Figure 3.4) but with a much stronger estimation noise, which likely originates from the noise in the spatial error variance difference estimates (Figure 3.5). However, the results of the analytical investigation of the optimum Kalman gain weights are in agreement with numerical results shown earlier in Section 3.3. That is, in the case where $\delta Q = \delta R$, negligible weight is given to the neighboring observations if $\Delta C \approx 0$. In the case of $\delta Q \neq \delta R$, no weight is given in the case of $\sqrt{\delta Q} - \sqrt{\delta R} = QR\Delta C\sqrt{\delta R}$). In these cases, the 1D and 2D filtering cases should converge. In areas where non-zero weights are present (i.e., where $\delta Q \neq \delta R$ and $C_Q > 0.7$ - 0.8 [-], or where $\Delta C \neq 0$), these weights are very sensitive to parameter estimation noise which in most cases equals or even outweighs the marginal accuracy improvement associated with the transition from 1D to 2D filtering. Therefore (3.1) and (3.2) provide an analytical explanation for the major qualitative results presented in Section 3.3. That

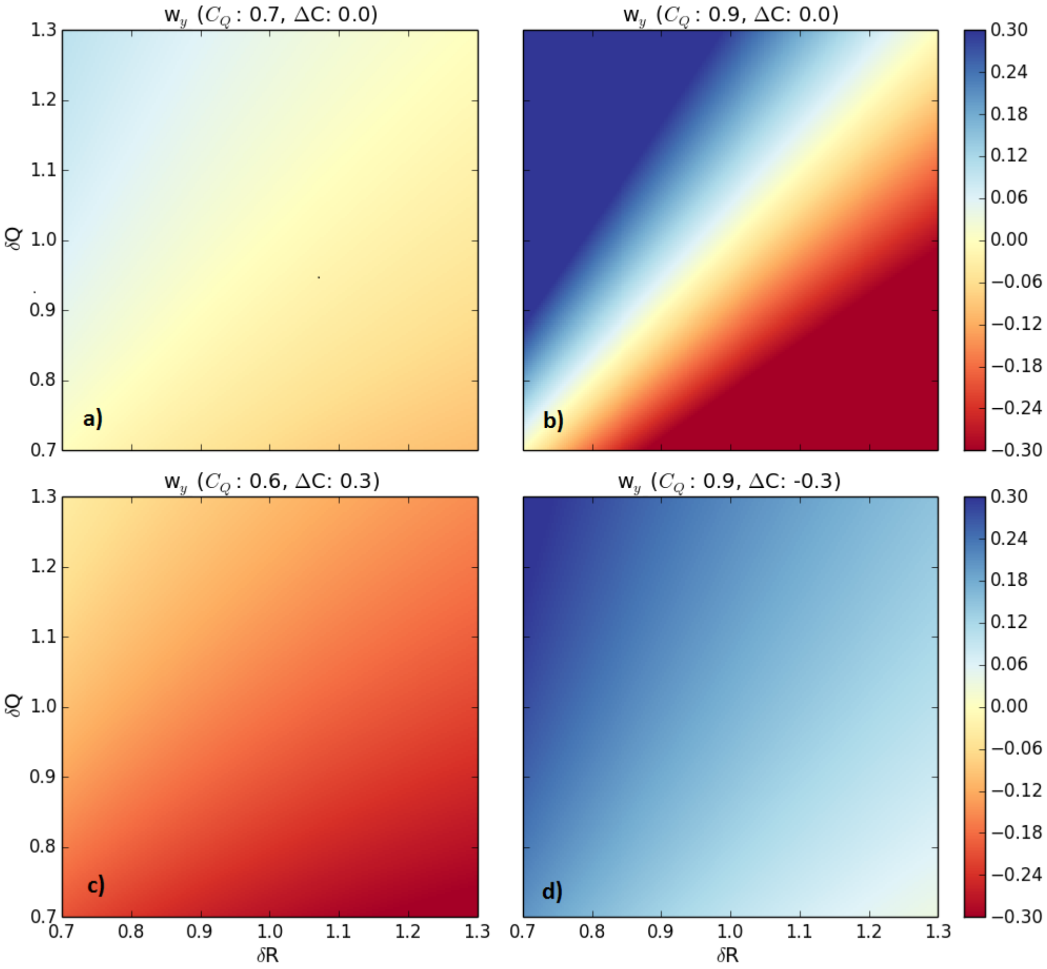

Figure 3.1: *Optimal Kalman gain weight (based on the simplified expression given in (3.1)) for observations in neighboring pixels as a function of δR (x-axis) and δQ (y-axis) for different C_Q and ΔC ((a)-(d)).*

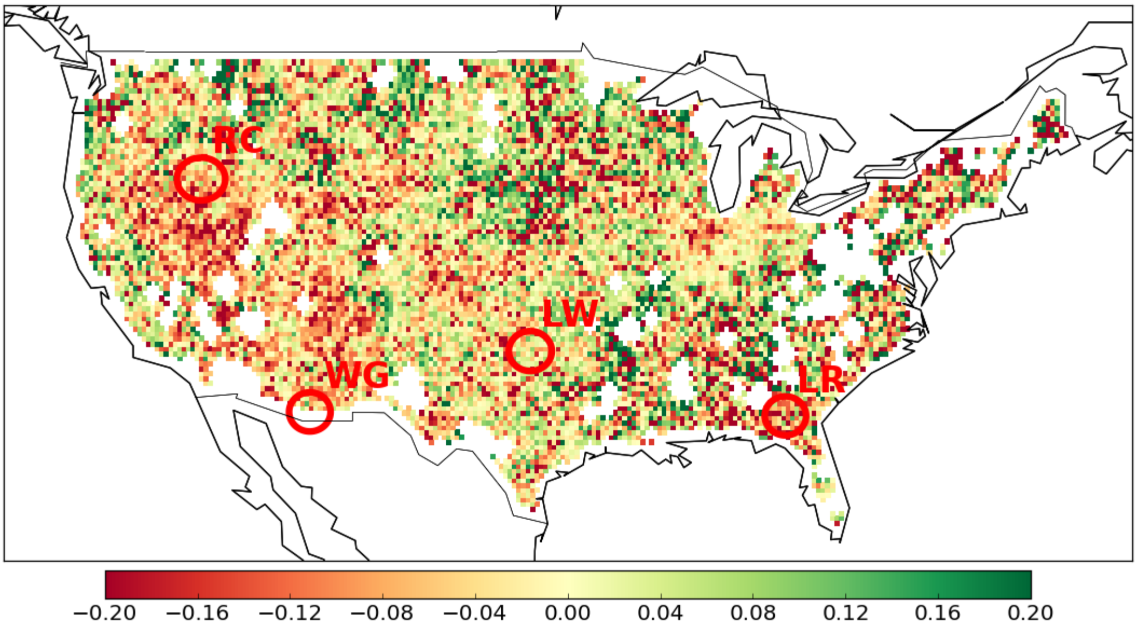

Figure 3.2: *Optimal Kalman gain weights for neighboring pixels calculated using the simplified expression in* (3.1). *Plotted values reflect an average across all cardinal directions and red circles mark the USDA ARS watershed locations.*

is, the dependency of 2D Kalman filtering performance on relative spatial error auto-correlation differences between the model and the observations and the lack of any performance increase when introducing a 2D filtering strategy.

3.5 Summary and conclusion

Land data assimilation systems represent the most common pathway by which remotely-sensed soil moisture products are integrated into existing climate, agricultural, and water resources applications. Such systems are usually applied using a 1D strategy where models are updated only with spatially-collocated observations. Nevertheless, if the errors in the model and/or the observation data sets are auto-correlated in space, soil moisture information can also be laterally propagated to neighboring modeling pixels by utilizing a 2D data assimilation strategy.

Information regarding the statistical structure of errors in remotely-sensed and modeled soil moisture is required to effectively parameterize data assimilation systems - particularly for such 2D cases. Here we develop and apply a triple collocation (TC) based approach to estimate statistical error parameters via the cross-comparison of active-microwave, passive-microwave, and model-based soil moisture estimates. In particular, a novel extension of TC is introduced which allows for the estimation of spatial error auto-correlation information.

In a synthetic experiment, the achieved RMSE of the error auto-correlation estimates was below 0.02 [-]. An application to real remotely-sensed (i.e., ASCAT-based) and modeled (i.e., TRMM-based) soil moisture datasets reveals that significant auto-correlations exist in all examined soil moisture products (Figure 3.4).

Such significant auto-correlation would seem to imply a large advantage associated with the application of a 2D data assimilation approach relative to a 1D baseline system. However, based on validation results against ground-based observations, no such advantage was observed (Figures 3.6 and 3.7). This lack of sensitivity is clarified via an analytical derivation of the (simplified) steady-state 2D Kalman filter gain (Section 3.4). The derived gain reveals that the optimum Kalman gain weight applied to neighboring pixels is highly dependent on the presence of differences between model and observation error variances and covariances rather than on their absolute values. If these differences are small, no weight is given to neighboring pixels, even if the absolute model and observation error auto-correlations are very high. Furthermore, even if non-zero, the estimated optimum weights are very sensitive to error parameter estimation noise. That is, even small variations in the error variance and/or spatial error auto-correlation estimates can lead to large fluctuations in the estimated optimal weights applied to neighboring pixels. This suggests that the application of a localized 2D Kalman filter to this particular assimilation problem will be relatively non-robust and highly-sensitive to relatively small variations in error parameters.

In a synthetic experiment, a significant skill gain of the 2D filter with respect to a 1D filter (in terms of correlation with the synthetic truth) is achieved for very large absolute differences between model and observation error auto-correlations ($|\Delta C| > 0.7$ - 0.8 [-]). However, for the well-controlled synthetic case presented here, the mean and maximum obtained correlation improvements over the entire ΔC range are 0.03 and 0.16, respectively. That is, the maximum achievable improvement of the 2D filter is only marginal.

For a real-data experiment, triple collocation was applied over the entire (contiguous) United States to investigate error variance and error auto-correlation differences between the model and the assimilated observations. Model versus observation error auto-correlation differences were in general rather low (Figure 3.4). Based on synthetic results presented in Figure 3.2, this would seem to suggest only modest advantages associated with the 2D assimilation of ASCAT observations into (3.1). To confirm this, an in situ evaluation of the filter performance was carried out over four heavily-equipped watershed sites as well as over 228 single ground stations, homogeneously distributed over the US, covering a large variety of possible weight levels. On average, the performance of the 2D filter (measured as the correlation with the in situ observations) is the same as that of the 1D baseline (ubRMSD = 0.05 [-]) with a very slight tendency of a performance decrease for the 2D case. No relationship between filter performance and average weight level of neighboring pixels is observed. A clear analytical justification for this lack of significant improvement is presented in Section 3.4.

These results are somewhat at odds with previous studies which identified greater levels of improvement associated with the transition to a 2D data assimilation scheme (e.g., *Reichle and Koster*, 2003; *Han et al.*, 2012). However, it should be noted that most of these previous studies were based on synthetic twin experiments which: 1) neglect the potential consequences of mis-parameterizing a data assimilation system (since error statistics are perfectly known) and 2) are based on (relatively) ad hoc assumptions concerning the spatial error characteristics of the model background and the assimilated soil moisture observations. This study improves on both of these points by characterizing errors using a TC strategy and then objectively evaluating the robustness of a real-data 2D KF implementation (based on this characterization) via comparison against independent observations. Our findings reveal that (commonly made) crude assumptions of spatial error statistics in a 2D system will at best maintain the performance of a 1D approach or - more likely - worsen the filter forecasts because an over- or underestimation of error auto-correlation difference can lead to an overestimation (in absolute terms) of the Kalman gain weight for the neighboring pixel.

However, it is worth noting that previous studies identified increased value in 2D filtering when compensating for permanent spatial gaps in the availability of soil moisture observations (*Han et al.*, 2012). It should be stressed that this particular case is not examined here and no updating is attempted for pixels consistently lacking remotely-sensed soil moisture retrievals (see e.g., masked areas in Figure 5). A fuller application considering this gap-filling potential may lead to a slightly-modified assessment of 2D soil moisture data assimilation.

Appendix

3.A Justification for using a simplified steady-state Kalman gain expression

The weights that are given to neighboring pixels are determined through the off-diagonal elements of the steady-state Kalman gain $\mathbf{K}$, which is obtained from (3.4)-(3.6) for $t \gg 0$. For simplicity, let us consider only one neighboring pixel for the assimilation, so that $\mathbf{K}$ reduces to a 2×2 matrix where the off-diagonal element is the direct weight for this pixel. Due to the temporal memory of the model, the steady-state Kalman gain has to be solved by combining (3.4)-(3.6) as:

$$\begin{aligned} \mathbf{K} &= \mathbf{M}\mathbf{H}^\top(\mathbf{H}\mathbf{M}\mathbf{H}^\top + \mathbf{R})^{-1} \\ \mathbf{M} &= \gamma[\mathbf{M} - \mathbf{M}\mathbf{H}^\top(\mathbf{H}\mathbf{M}\mathbf{H}^\top + \mathbf{R})^{-1}\mathbf{H}\mathbf{M}]\,\gamma^\top + \mathbf{Q} \end{aligned} \tag{3.A.1}$$

However, the analytical steady-state solution of this quadratic matrix equation is mathematically demanding and produces a complex, and difficult to interpret, analytical expression (even for low-dimensional cases).

For investigating the impact of *spatial* error auto-covariances on the weight of the neighboring pixel we can simplify things by neglecting the *temporal* model memory (i.e., assuming $\gamma = 0$ for the auto-regressive model in (3.1)) so that $\mathbf{M}_t = \mathbf{Q}$. This leads to a greatly simplified (and much more readily interpretable) expression for the steady-state Kalman gain, given as:

$$\mathbf{K} = \mathbf{Q}\mathbf{H}^\top(\mathbf{H}\mathbf{Q}\mathbf{H}^\top + \mathbf{R})^{-1} \tag{3.A.2}$$

whose off-diagonal element is the simplified weight given in (3.1). This simplified weight deviates from the full weight through a (slightly non-linear) scaling function. Due to high mathematical complexity of the analytical solution of (3.A.1) we will omit its derivation and calculate it numerically by iterating over (3.4)-(3.6) until $t \gg 0$. Figure 3.A.1 shows a comparison between the thereby obtained full weights and the simplified weights obtained directly from (3.1) - calculated for all 16245 synthetic data sets described in Section 3.3.1. One can see an almost perfect, slightly non-linear correlation ($\rho_p = 0.96$; $\rho_s = 1.00$), which is valid for a wide range of values. Note that also the majority of weights estimated in the real data experiment fall within this valid range (Figure 3.7). Therefore, the use of the simplified analytical expression in (3.1) for these weights (in Section 3.4) appears to be well-justified.

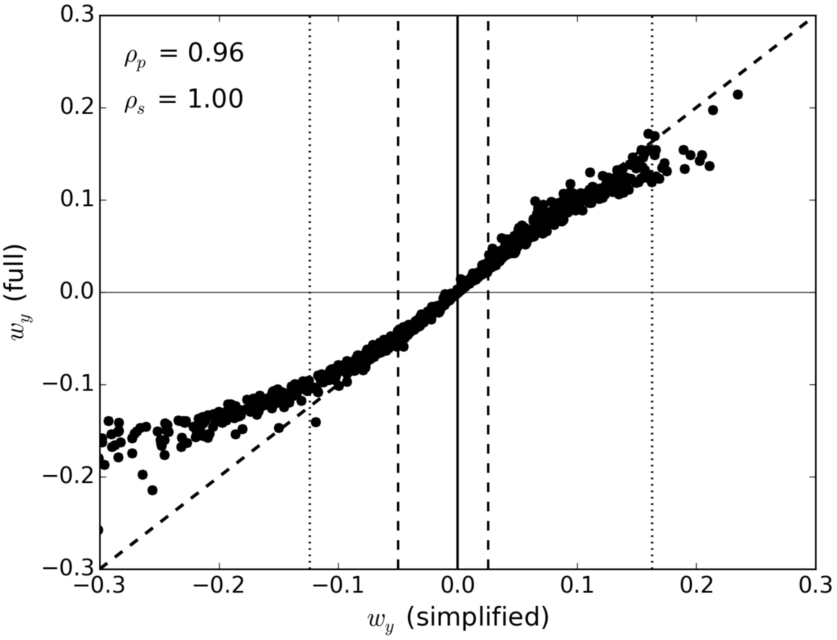

Figure 3.A.1: *Scatterplot of optimal weights (for observations in neighboring pixels) derived using the simplified analytical expression in (3.1) vs. full weights calculated via the numerical iteration of the full Kalman filter for all 16245 synthetic data sets. ρ_p is the Pearson- and ρ_s the Spearman correlation coefficient. Vertical lines represent the 5% and 95% quantile (dotted), the upper and the lower quartile (dashed), and the median (solid) of full optimal Kalman gain weights calculated at the 228 SCAN and USCRN sites (Figure 3.7).*

Chapter 4

Estimating error cross-correlations in soil moisture data sets using extended collocation analysis

Global soil moisture records are essential for studying the role of hydrologic processes within the larger earth system. Various studies have shown the benefit of assimilating satellite-based soil moisture data into water balance models or merging multi-source soil moisture retrievals into a unified data set. However, this requires an appropriate parameterization of the error structures of the underlying data sets. While triple collocation (TC) analysis has been widely recognized as a powerful tool for estimating random error variances of coarse-resolution soil moisture data sets, the estimation of error cross covariances remains an unresolved challenge. Here we propose a method — referred to as extended collocation (EC) analysis — for estimating error cross-correlations by generalizing the TC method to an arbitrary number of data sets and relaxing the therein made assumption of zero error cross-correlation for certain data set combinations. A synthetic experiment shows that EC analysis is able to reliably recover true error cross-correlation levels. Applied to real soil moisture retrievals from Advanced Microwave Scanning Radiometer-EOS (AMSR-E) C-band and X-band observations together with advanced scatterometer (ASCAT) retrievals, modeled data from Global Land Data Assimilation System (GLDAS)-Noah and in situ measurements drawn from the International Soil Moisture Network, EC yields reasonable and strong nonzero error cross-correlations between the two AMSR-E products. Against expectation, nonzero error cross-correlations are also found between ASCAT and AMSR-E. We conclude that the proposed EC method represents an important step toward a fully parameterized error covariance matrix for coarse-resolution soil moisture data sets, which is vital for any rigorous data assimilation framework or data merging scheme.

---

* This chapter is an edited version of: 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.

4.1 Introduction

Consistent global soil moisture records are essential for studying hydrology driven phenomena of the Earth system such as climate change, vegetation growth, and many others (*Legates et al.*, 2011). Various studies have shown the benefit of blending satellite-based soil moisture observations from multiple platforms into a unified data set (*Liu et al.*, 2011b, 2012) or assimilating them into water balance models in order to generate a continuous merged (model/remote-sensing) soil moisture analysis product (*Bolten and Crow*, 2012; *de Rosnay et al.*, 2013). However, such merging and assimilation frameworks require an appropriate statistical parameterization of the error structures of both the land surface model and the remote sensing data, which is often difficult to obtain in practice. This error parameterization problem becomes even more challenging if errors between different input data sets are correlated as this requires the parameterization of error covariances (i.e., the off-diagonal elements of the error covariance matrix) in addition to error variances (i.e., the diagonal elements of the error covariance matrix).

In the past, off-diagonal elements in the error covariance matrix were commonly neglected as there was no method available for reliably estimating these elements (*Yilmaz et al.*, 2012). At the same time, the increasing simultaneous availability of various active and passive satellite-based sensors (e.g., ASCAT onboard MetOp-A and MetOp-B, SMAP, SMOS, AMSR-E, AMSR2, etc.) inevitably leads to the need for a fully parameterized error covariance matrix, which is vital for any statistically rigorous attempt to merge multi-source soil moisture retrievals into a unified data set (*Crow et al.*, 2015).

Triple collocation (TC) analysis (*Stoffelen*, 1998) has been widely recognized as a powerful tool for parameterizing the diagonal elements of the error covariance matrix (*Crow and Van den Berg*, 2010). A first attempt to additionally estimate off-diagonal elements of the error covariance matrix was made by *Crow and Yilmaz* (2014) who analytically combined TC analysis with Kalman filter innovation analysis - referred to as Auto-Tuned Land Data Assimilation System (ATLAS) - yet the stability of the thereby obtained error cross-covariance estimates has not been proven over larger scales. More recently, *Crow et al.* (2015) proposed a TC-based approach to estimate off-diagonal elements by using lagged variables (i.e., temporally shifted representations of a particular data set; *Su et al.*, 2014a) to generate data set triplets with uncorrelated errors, which can also provide consistent error variance estimates. Subtracting these estimates from error variance estimates obtained from a triplet using the corresponding data set together with two data sets that have correlated errors then yields an estimate of their error covariance. However, error cross-covariance estimates produced by this technique can become biased in the presence of temporal error auto-correlation (*Crow et al.*, 2015). Another extension of TC that also tolerates the existence of non-zero error cross-correlations when using more than three data sets for the collocation was proposed by (*Pan et al.*, 2015). It solves the collocation problem through Pythagorean constraints in Hilbert space, yet it does not yield estimates for non-zero error cross-correlations. Instead, it splits all considered data sets into so-called structural groups, whithin which the data sets are likely to have correlated errors. Random error variances of each data set in each group are then estimated as two components: One part that is correlated with the errors of the other data sets (within the same group), and the remaining part that is entirely independent from all other data sets (within all groups). Summing these two components up yields estimates for the individual total error varaince of all data sets.

Here we propose an alternative method for estimating error cross-correlations by generalizing TC analysis to an arbitrary number of $N > 3$ data sets following *Zwieback et al.* (2012) and relaxing the assumption of zero error-cross correlation for a limited number of data set combinations. The resulting method is referred to as extended collocation (EC) analysis and allows for the estimation of a limited number of non-zero error cross-correlations - in addition to error variance and scaling coefficient estimates for all considered data sets - depending on the number of data sets used and their assumed underlying error structure. Of particular importance will be the estimation of error cross-correlation amongst different active-satellite-based data sets (e.g., MetOp-A and MetOp-B ASCAT), amongst passive-satellite-based data sets (e.g., SMOS, AMSR2 and WindSat), amongst data sets derived from the same sensor using different retrieval algorithms (e.g., SMOS L3, SMOS LPRM), and amongst land surface models with similar atmospheric forcing (e.g., ERA-Land and GLDAS-Noah), all of which are simultaneously resolvable in the EC analysis framework.

For simplicity and without any loss of generality, the method will be discussed and demonstrated using maximum five data sets. Note that *Pierdicca et al.* (2015) recently proposed to extend TC analysis with a fourth data set and to solve this quadruple collocation (QC) problem as an overdetermined system of three possible triplets in a least-squares sense. This minimizes the uncertainty of the individual error estimates, but still requires uncorrelated errors between all four data sets. For the EC method proposed here we follow *Pierdicca et al.* (2015) in solving the collocation system of equations in a least-squares sense in cases where the system remains overdetermined after additionally leveraging some degrees of freedom to estimate further parameters (i.e., error cross-correlations). It is worth mentioning that even though only soil moisture data sets are considered in this study, EC is - just like TC - also applicable to other geophysical variables in hydrometeorology and oceanography (e.g., *Vogelzang et al.*, 2011; *Caires and Sterl*, 2003; *Roebeling et al.*, 2012; *Fang et al.*, 2012).

The method will be derived in Section 4.2. Section 4.3 shows an evaluation of the method using both synthetic identical twin experiments and a real data experiment.

4.2 Background

Our proposed EC method is a generalization of the well-known triple collocation (TC) analysis (*Stoffelen*, 1998), which is commonly used for estimating the individual error variances of three spatio-temporally collocated soil moisture data sets with mutually uncorrelated random errors (*Scipal et al.*, 2008; *Dorigo et al.*, 2010). In the following sections we will derive the estimators using the so-called covariance notation for the collocation problem (*Stoffelen*, 1998; *Su et al.*, 2014b; *Gruber et al.*, 2015).

4.2.1 Triple collocation

Classical TC analysis assumes a linear error model of the following form:

$$i = \alpha_i + \beta_i \Theta + \varepsilon_i \tag{4.1}$$

with $i \in [a, b, c]$ representing three spatially and temporally collocated soil moisture data sets; Θ is the true soil moisture state; α_i and β_i are additive and multiplicative biases in data set i; and ε_i is zero-mean random noise. By using the error model in (4.1), the data set variances and covariances can be written as:

$$\begin{aligned} \sigma_i^2 &= \beta_i^2 \sigma_\Theta^2 + 2\beta_i \sigma_{\Theta\varepsilon_i} + \sigma_{\varepsilon_i}^2 \\ \sigma_{ij} &= \beta_i \beta_j \sigma_\Theta^2 + \beta_j \sigma_{\Theta\varepsilon_i} + \beta_i \sigma_{\Theta\varepsilon_j} + \sigma_{\varepsilon_i \varepsilon_j} \end{aligned} \tag{4.2}$$

with $i, j \in [a, b, c]$. TC analysis assumes error orthogonality ($\sigma_{\Theta\varepsilon_i} = 0$), and zero error cross-correlation ($\sigma_{\varepsilon_i\varepsilon_j} = 0$ for $i \neq j$). (4.2) thus simplifies to:

$$\begin{aligned} \sigma_i^2 &= \beta_i^2 \sigma_\Theta^2 + \sigma_{\varepsilon_i}^2 \\ \sigma_{ij} &= \beta_i \beta_j \sigma_\Theta^2 \end{aligned} \tag{4.3}$$

From (4.3) we can now derive direct estimates for both the soil moisture signal variances ($\beta_i^2 \sigma_\Theta^2$) and the random error variances ($\sigma_{\varepsilon_i}^2$) of the individual data sets as:

$$\beta_i^2 \sigma_\Theta^2 = \frac{\sigma_{ij}\sigma_{ik}}{\sigma_{jk}} \qquad \sigma_{\varepsilon_i}^2 = \sigma_i^2 - \frac{\sigma_{ij}\sigma_{ik}}{\sigma_{jk}} \tag{4.4}$$

with $i, j, k \in [a, b, c]$ and $i \neq j \neq k$. These are the final error estimates obtained from TC analysis, which allow for either a direct investigation of the error variances ($\sigma_{\varepsilon_i}^2$), or for an investigation of the signal-to-noise ratios ($\mathrm{SNR}_i = \frac{\beta_i^2 \sigma_\Theta^2}{\sigma_{\varepsilon_i}^2}$) of the data sets. However, these estimates become biased in the presence of non-zero error cross-correlations and/or non-orthogonal errors, as it can be seen in (4.2). The first quantitative investigation of such biases due to violations in TC assumptions was recently made by Yilmaz and Crow (2014). While results of this study suggest that both non-zero error cross-correlation and non-orthogonal errors may exist in typical soil moisture data sets, it was also found that the impact of error cross-correlations are of greater importance than that of error non-orthogonalities. This is because the impact of the latter can be dampened or even compensated if their magnitude is approximately equal for all data sets, and also because errors of different data sets that are non-orthogonal are typically also cross-correlated.

4.2.2 Extended collocation problem

Let us now generalize the TC problem in (4.3) for an arbitrary number of N data sets (Zwieback et al., 2012) and relax the assumption of zero error cross-correlation for some data set combinations while maintaining the assumption of orthogonal errors for all data sets. According to (4.2), the data set covariances then write as:

$$\sigma_{ij} = \begin{cases} \beta_i \beta_j \sigma_\Theta^2 & \forall\ i, j \quad \text{where} \quad \sigma_{\varepsilon_i \varepsilon_j} = 0 \\ \beta_i \beta_j \sigma_\Theta^2 + \sigma_{\varepsilon_i \varepsilon_j} & \forall\ i, j \quad \text{where} \quad \sigma_{\varepsilon_i \varepsilon_j} \neq 0 \end{cases} \tag{4.5}$$

with $i \neq j$. Cross-covariances between errors in i and j can then be directly estimated from (4.5) as:

$$\beta_i \beta_j \sigma_\Theta^2 = \frac{\sigma_{ik}\sigma_{jl}}{\sigma_{kl}} \qquad\qquad \sigma_{\varepsilon_i \varepsilon_j} = \sigma_{ij} - \frac{\sigma_{ik}\sigma_{jl}}{\sigma_{kl}} \tag{4.6}$$

with $i \neq j \neq k \neq l$ where $\sigma_{\varepsilon_i \varepsilon_k}$, $\sigma_{\varepsilon_j \varepsilon_l}$, and $\sigma_{\varepsilon_k \varepsilon_l}$ are required to be zero. Error cross-correlations can be further derived by simply dividing (4.6) through the error standard deviations obtained using (4.4) applied on data set triplets with mutually uncorrelated errors, provided that they are available (see Section 4.2.4).

Notice that (4.6) uses a combination of exactly four different covariances (between four data sets pairs), three of which are required to have uncorrelated errors. However, the availability of four data sets already provides six possible data set pairs (i.e., six different covariances), increasing with the number of data set sets (N) as $\frac{N!}{2(N-2)!}$. Therefore, we can typically define a certain number of redundant estimators for $\sigma_{\varepsilon_i \varepsilon_j}$. The same holds for the signal- and error variance estimates, i.e., for the $\beta_i^2 \sigma_\Theta^2$ and $\sigma_{\varepsilon_i}^2$ obtained from (4.4), which require three data set pairs, all of which must have uncorrelated errors. This redundancy allows us to solve the EC problem in a least squares sense in order to reduce estimation uncertainties in the error variance and -covariance estimates (Su et al., 2014a; Pierdicca et al., 2015).

4.2.3 Least-squares solution

Let us therefore summarize the final collocation system of equations as:

$$
\begin{array}{lll}
\sigma_i^2 = \beta_i^2\sigma_\Theta^2 + \sigma_{\varepsilon_i}^2 & \forall\, i & \\
\sigma_{ij} = \beta_i\beta_j\sigma_\Theta^2 + \sigma_{\varepsilon_i\varepsilon_j} & \forall\, i,j \quad \text{where} & \sigma_{\varepsilon_i\varepsilon_j} \neq 0 \\
\dfrac{\sigma_{ij}\sigma_{ik}}{\sigma_{jk}} = \beta_i^2\sigma_\Theta^2 & \forall\, i,j,k \quad \text{where} & \sigma_{\varepsilon_i\varepsilon_j} = \sigma_{\varepsilon_i\varepsilon_k} = \sigma_{\varepsilon_j\varepsilon_k} = 0 \\
\dfrac{\sigma_{ik}\sigma_{jl}}{\sigma_{kl}} = \beta_i\beta_j\sigma_\Theta^2 & \forall\, i,j,k,l \quad \text{where} & \sigma_{\varepsilon_i\varepsilon_k} = \sigma_{\varepsilon_j\varepsilon_l} = \sigma_{\varepsilon_k\varepsilon_l} = 0
\end{array}
\tag{4.7}
$$

In matrix notation (4.7) writes as:

$$
\mathbf{y} = \begin{pmatrix} \sigma_i^2 \\ \sigma_{ij} \\ \frac{\sigma_{ij}\sigma_{ik}}{\sigma_{jk}} \\ \frac{\sigma_{ij}\sigma_{kl}}{\sigma_{jl}} \end{pmatrix}
\qquad
\mathbf{A} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}
\qquad
\mathbf{x} = \begin{pmatrix} \beta_i^2\sigma_\Theta^2 \\ \beta_i\beta_j\sigma_\Theta^2 \\ \sigma_{\varepsilon_i}^2 \\ \sigma_{\varepsilon_i\varepsilon_j} \end{pmatrix}
\tag{4.8}
$$

where $\mathbf{y} = \mathbf{Ax}$; $\mathbf{y}$ is the (known) observation vector; $\mathbf{A}$ is the design matrix, and $\mathbf{x}$ is the vector of unknown parameters. The actual dimensions of $\mathbf{y}$, $\mathbf{A}$, and $\mathbf{x}$ depend on the number of data sets used and on the number of data set pairs which are (a-priori) assumed to have correlated errors. This also determines the degree of redundancy in $\mathbf{Ax}$. As an example, for the case of four data sets - referred to as the quadruple collocation (QC) scenario with $i, j, k, l \in [a, b, c, d]$ - with only a and b having correlated errors ($\sigma_{\varepsilon_a\varepsilon_b} \neq 0$), (4.8) takes the form:

$$
\mathbf{y} = \begin{pmatrix} \sigma_a^2 \\ \sigma_b^2 \\ \sigma_c^2 \\ \sigma_d^2 \\ \sigma_{ab} \\ \frac{\sigma_{ac}\sigma_{ad}}{\sigma_{cd}} \\ \frac{\sigma_{bc}\sigma_{bd}}{\sigma_{cd}} \\ \frac{\sigma_{ac}\sigma_{cd}}{\sigma_{ad}} \\ \frac{\sigma_{bc}\sigma_{cd}}{\sigma_{bd}} \\ \frac{\sigma_{ad}\sigma_{cd}}{\sigma_{ac}} \\ \frac{\sigma_{bd}\sigma_{cd}}{\sigma_{bc}} \\ \frac{\sigma_{ac}\sigma_{bd}}{\sigma_{cd}} \\ \frac{\sigma_{ad}\sigma_{bc}}{\sigma_{cd}} \end{pmatrix}
\qquad
\mathbf{A} = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}
\qquad
\mathbf{x} = \begin{pmatrix} \beta_a^2\sigma_\Theta^2 \\ \beta_b^2\sigma_\Theta^2 \\ \beta_c^2\sigma_\Theta^2 \\ \beta_d^2\sigma_\Theta^2 \\ \beta_a\beta_b\sigma_\Theta^2 \\ \sigma_{\varepsilon_a}^2 \\ \sigma_{\varepsilon_b}^2 \\ \sigma_{\varepsilon_c}^2 \\ \sigma_{\varepsilon_d}^2 \\ \sigma_{\varepsilon_a\varepsilon_b} \end{pmatrix}
\tag{4.9}
$$

The least-squares solution for the parameters $\mathbf{x}$ is then given as:

$$\hat{\mathbf{x}} = \left(\mathbf{A}^\top\mathbf{A}\right)^{-1}\mathbf{A}^\top\mathbf{y} \tag{4.10}$$

Notice that the QC case ($N = 4$) with only one non-zero error cross-correlation was chosen merely as an example for demonstration purposes. (4.9) can be easily extended to any number of $N > 4$ data sets, which allows also for the estimation of more than one non-zero error cross-correlation, for example between multiple active satellite-based and multiple passive satellite-based soil moisture data sets. However, regardless of the number of data sets used in EC analysis, not every possible error structure is resolvable.

4.2.4 Resolvable error structures

In (4.6) we see that the consistency of the error cross-covariance estimator requires zero error cross-covariance between some specific data set combinations, i.e., $\sigma_{\varepsilon_i\varepsilon_k} = \sigma_{\varepsilon_j\varepsilon_l} = \sigma_{\varepsilon_k\varepsilon_l} = 0$. The same holds for the signal- and error variance estimators in (4.4), which require $\sigma_{\varepsilon_i\varepsilon_j}$, $\sigma_{\varepsilon_i\varepsilon_k}$, and $\sigma_{\varepsilon_j\varepsilon_k}$ to be zero. If any of these were allowed to be non-zero, the matrix $(\mathbf{A}^\top\mathbf{A})$ would become singular and the collocation system of equations in (4.10) cannot be solved. However, regardless of the number of data sets used we can define the requirement on the invertability of the the matrix $(\mathbf{A}^\top\mathbf{A})$ as follows: Each member of the data set pairs with cross-correlated errors must also be a member of at least one data set triplet with mutually uncorrelated errors. For example, when using two passive satellite-based data sets - which are those assumed to have correlated errors - together with one active satellite-based and one modelled data set, we can define two triplets comprised of the active microwave based, the modelled, and one passive satellite-based data set, respectively, both of which have fully independent error structures. In this case, $(\mathbf{A}^\top\mathbf{A})$ can be inverted, and the collocation system of equations can be solved. More generally speaking, $\mathbf{A}^\top\mathbf{A}$ has to have full rank. Therefore, the rank of $\mathbf{A}^\top\mathbf{A}$ (and thus also of $\mathbf{A}$) has to be equal to the size of $\mathbf{x}$.

4.3 Demonstration

In the following sections we will evaluate the EC method using both synthetic identical twin experiments and a real data analysis. For simplicity and without any loss of generality we will limit the demonstration to scenarios where either four or five data sets are available.

4.3.1 Synthetic experiment

For the synthetic experiment we limit the number of data sets (N) to $N = 4$ (i.e., to the QC scenario) with only one data set pair having cross-correlated errors. This represents the worst case (in the synthetic case) since the inclusion of more data sets would increase the degrees of freedom in the collocation system of equations, which would lead to an increased precision of the estimates.

A true soil moisture reference Θ is first generated via an unperturbed integration of the Antecedent Precipitation Index (API) model ($\Theta_t = \gamma \Theta_{t-1} + P_t$; where t is the time index, the loss variable γ is held fixed at 0.85, and the precipitation P is modelled as a Possion process (*Crow et al.*, 2012a)). Four soil moisture data sets are then generated by artificially perturbing the soil moisture reference with random noise containing varying cross-correlations, drawn from a multivariate normal distribution.

Synthetic soil moisture quadruplets are generated for a large number of different cases. Error cross-correlation levels between two of the data sets are systematically varied between 0.0 [-] and 1.0 [-] in increments of 0.1 [-], and error variance levels are varied in all four data sets between 40 mm^2 and 600 mm^2 in increments of 80 mm^2, which corresponds to a SNR between about -6 dB and $+6$ dB, which is a typical range for soil moisture data sets (*Gruber et al.*, 2015). Altogether, this requires the generation of 45056 separate synthetic data sets. The sample size of each data set is 750 days which is approximately the average sample size that is available for the real data experiment (see Section 4.3.2). The EC based error cross-correlation estimates for these 45056 data sets - obtained using (4.10) - are shown in Figure 4.1. True error cross-correlation levels can be recovered without bias and with negligible RMSE (0.08 [-]), which decreases with increasing error cross-correlation. Therefore, the application of EC for accurately estimating error cross-correlations appears plausible. Note that the apparent increase in estimation accuracy with increasing error cross-correlation magnitude originates from the non-linear $(\cdots)^{-1}$ dependency on error variance estimates when converting the error cross-covariance estimates to error cross-correlations. The uncertainties of the error cross-covariance estimates alone do not show such a dependence.

4.3.2 Real data experiment

In this section we further evaluate the EC method by applying it to real data. The soil moisture data sets used for this study are: (i) passive satellite-based retrievals from the AMSR-E C-band channel, (ii) passive satellite-based retrievals from the AMSR-E X-band channel, (iii) active satellite-based retrievals from ASCAT, (iv) soil moisture estimates from the GLDAS-Noah land surface model, and (v) ground measurements from globally-distributed in situ stations drawn from the International Soil Moisture Network.

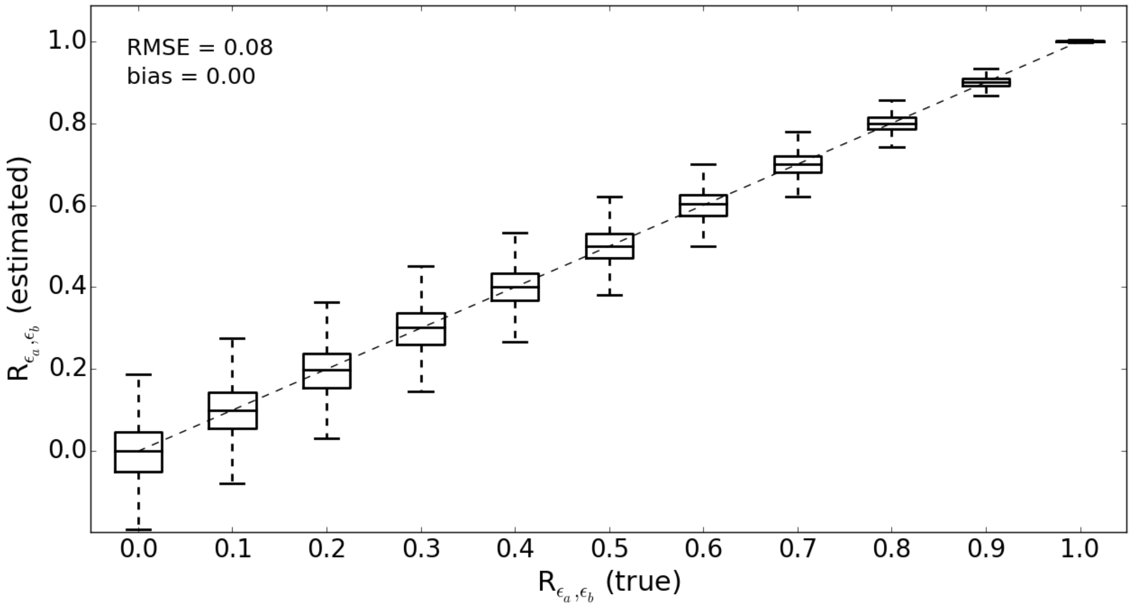

Figure 4.1: *The median and the inter-quartile-range (IQR) of estimated error cross-correlations (y-axis) at different true cross-correlation levels (x-axis) and different error variance levels. Whiskers represent 1.5 times the IQR. The estimation bias and RMSE averaged over all samples are provided on the top left hand corner. The sample size is 750 days.*

While active-based, passive-based, modelled, and in situ soil moisture estimates are widely-assumed to have mutually independent error structures, the two AMSR-E data sets from two different frequency channels are very likely to have significant non-zero error cross-correlation due to instrumental and algorithmic identity. Here we use EC analysis to estimate these supposed error cross-correlations between multi-frequency AMSR-E retrievals and further test the assumption of zero error cross-correlation between AMSR-E and ASCAT retrievals.

Soil moisture estimates from AMSR-E are retrieved using the Land Parameter Retrieval Model (LPRM) Version 5 (*Owe et al.*, 2008) and provided by the VU University Amsterdam (VUA). Data is provided in volumetric units on a regular grid with 0.25 degrees grid spacing. Vegetation Optical Depth (VOD) estimates are used to filter out retrievals with a high uncertainty due to dense vegetation (*Parinussa et al.*, 2011). Usually, Radio Frequency Interference (RFI) estimates are used to switch from C- to X-band retrievals in RFI-contaminated areas (*Owe et al.*, 2008). Here we consider both C- and X-band retrievals separately in order to estimate their mutual error cross-correlation. RFI estimates are used to mask out areas with high contamination in either of the frequency bands.

The active satellite-based soil moisture data set is the H-25 SM-OBS-4 MetOp-A ASCAT time series product, retrieved using the TU Wien algorithm version WARP 5.5 R2.2 (*Wagner et al.*, 1999; *Naeimi*, 2009). ASCAT operates at C-band, retrieved soil moisture estimates are provided as degree of saturation at a spatial resolution of 25 km, regridded to a 12.5 km Discrete Global Grid (DGG).

The WARP Surface State Flag (SSF; *Naeimi et al.*, 2012) is used to remove measurements taken under frozen or freezing/thawing conditions.

The Global Land Data Assimilation System (GLDAS-) Noah model provides soil moisture data for four different depth layers at a spatial resolution of approximately 0.25 degrees in a 3-hourly sampling rate (*Rodell et al.*, 2004). Only the top layer (0-10 cm) is used in this study.

In situ data is drawn from the International Soil Moisture Network (ISMN), which is a data hosting facility that collects and harmonizes data from networks and field validation campaigns world-wide, and makes them available to the users on a centralized web platform (*Dorigo et al.*, 2011a,b). For this study we consider all stations that lie within the temporally overlapping period of ASCAT and AMSR-E, i.e., January 2007 - October 2011. Measurements from sensors which are placed deeper than 10 cm below the surface are excluded. The ISMN also flags suspicious measurements such as spikes or signal saturation as well as measurements taken under frozen conditions or exceeding physically meaningful value ranges, based on automated quality control procedures (*Dorigo et al.*, 2013). Measurements flagged as suspicious are excluded in this study. Data sets that meet the above described requirements are provided by the networks: AMMA-CATCH (*Pellarin et al.*, 2009), ARM (http://www.arm.gov/), COSMOS (*Zreda et al.*, 2008), GTK, HOBE (*Bircher et al.*, 2012), ICN (*Hollinger and Isard*, 1994), MAQU (*Su et al.*, 2011), MOL-RAO (http://www.dwd.de/mol/), OZNET (*Smith et al.*, 2012), PBO-H2O (*Larson et al.*, 2008), REMEDHUS (http://campus.usal.es/˜hidrus/), SASMAS (*Young et al.*, 2008), SCAN (http://www.wcc.nrcs.usda.gov/), SMOSMANIA (*Albergel et al.*, 2008), SNOTEL (*Leavesley et al.*, 2008), SWEX-POLAND (*Marczewski et al.*, 2010), UDC-SMOS (*Schlenz et al.*, 2012), UMBRIA (*Brocca et al.*, 2011), USCRN (*Bell et al.*, 2013), and USDA-ARS (*Jackson et al.*, 2010).

4.3.2.1 EC analysis over the ISMN

As mentioned in Section 4.2.4, EC requires at least two data sets whose errors are fully independent from the errors of all other data sets in addition to the data sets with assumed non-zero error cross-correlation. Therefore, both modelled and in situ data need to be included in the EC analysis when assuming non-zero error cross-correlations between ASCAT and AMSR-E. However, this results in spatially incomplete estimates due to the limited global coverage of available ground stations.

Figure 4.2 shows the error cross-correlation statistics between retrievals from the two AMSR-E channels, between ASCAT and AMSR-E C-band retrievals, and between ASCAT and AMSR-E X-band retrievals, respectively, for both absolute values (median: 0.82 / 0.27 / 0.25) and anomalies (median: 0.78 / 0.21 / 0.20) for all available stations. Anomalies were calculated by subtracting a five week moving-average window based climatology. Figures 4.3 and 4.4 further show the spatial distribution of error cross-correlation over regions with a higher station coverage, i.e., the Contiguous United States, Europe, and New South Wales (Australia) for absolute measurements

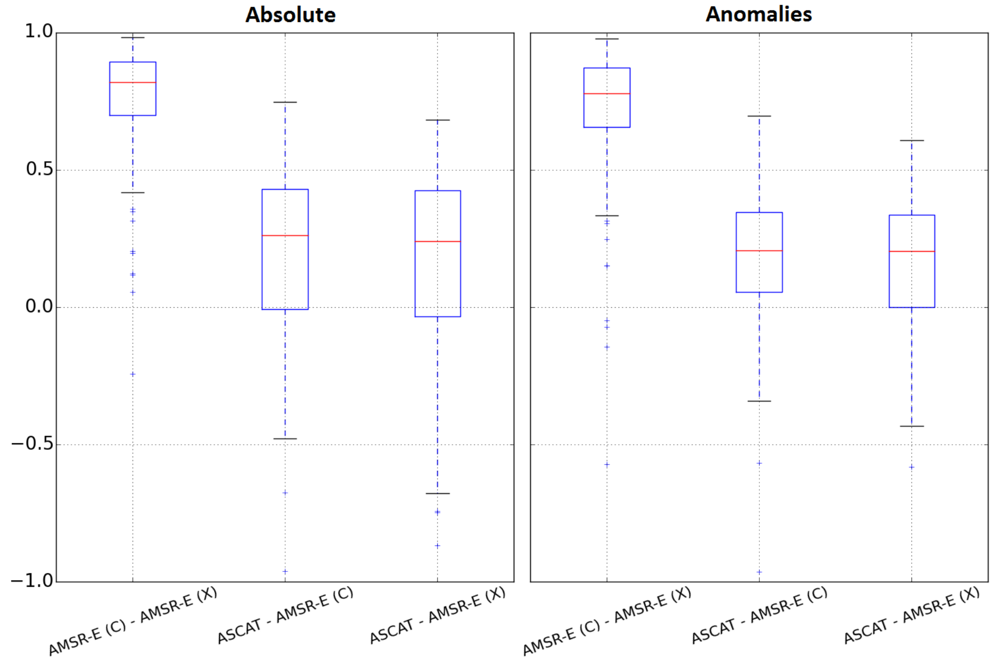

Figure 4.2: *The median and inter-quartile-range (IQR) of error cross-correlation estimates for soil moisture retrievals obtained from: the AMSR-E C- and X-band channels, ASCAT and the AMSR-E C-band channel, and ASCAT and the AMSR-E X-band channel for both absolute values (left) and anomalies (right). Whiskers represent 1.5 times the IQR. The sample size is 283 stations.*

and anomalies, respectively. As expected, cross-correlations between the errors of the AMSR-E data sets are very high in almost all regions. A detailed discussion on AMSR-E error cross-correlation will be provided later in Section 4.3.2.3. Against expectation, non-zero error cross-correlations exist - even though much lower - also between ASCAT and both AMSR-E frequency channels. These are slightly higher for absolute soil moisture retrievals than for anomalies and show some distinct spatial patterns: higher error cross-correlations over the Western US, which are more pronounced for absolute values than for anomalies, higher cross-correlations between errors of absolute values over the Mississippi region, which are not present in the anomalies, and higher values over agricultural areas in Australia for both absolute values and anomalies.

Most of the observed non-zero error cross-correlations seem to be located in areas where in situ stations typically have a limited spatial representativeness, for instance in the Western US where the topographic complexity is very high, or in the heavily irrigated Mississippi region. Therefore, the question arises whether the observed error cross-correlations in these regions are artificial biases due to limited representativeness of the ground measurements. In classical TC analysis, limited spatial representativeness causes a bias in the error variance estimates of the ground measurements, i.e., TC assigns them an additional representativeness error term (*Vogelzang and Stoffelen*, 2012; *Miralles et al.*, 2010; *Crow et al.*, 2012b; *Gruber et al.*, 2013, 2015). The error variance estimates of the coarse resolution data sets, on the other hand, remain unbiased. In the following section

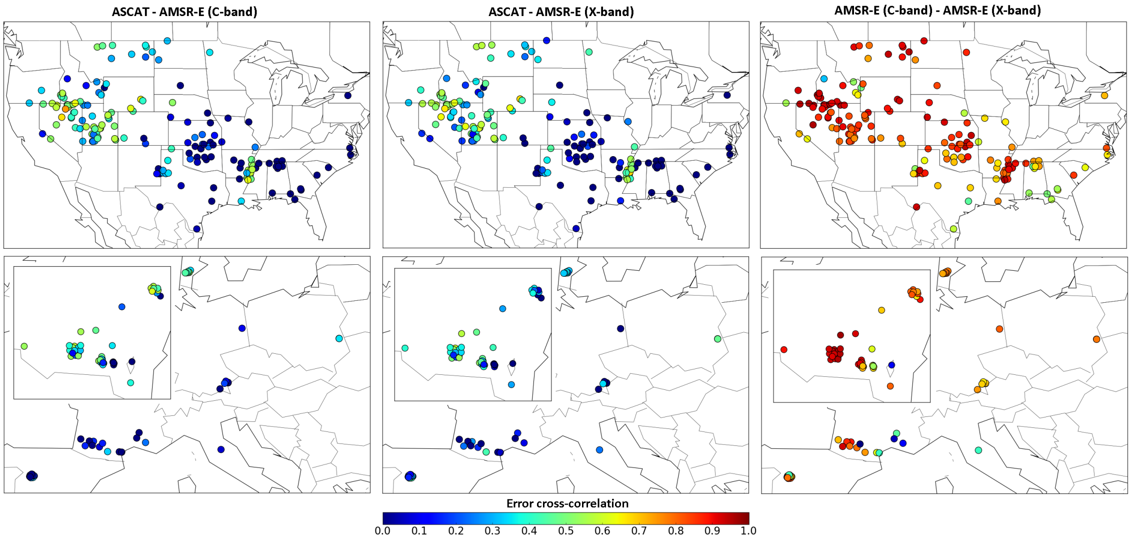

Figure 4.3: *Estimated cross-correlations between the errors in absolute soil moisture retrievals obtained from: ASCAT and the AMSR-E C-band channel (left), ASCAT and the AMSR-E X-band channel (middle), and the AMSR-E C- and X-band channels (right) for the Contiguous United States, Europe, and New South Wales (Australia).*

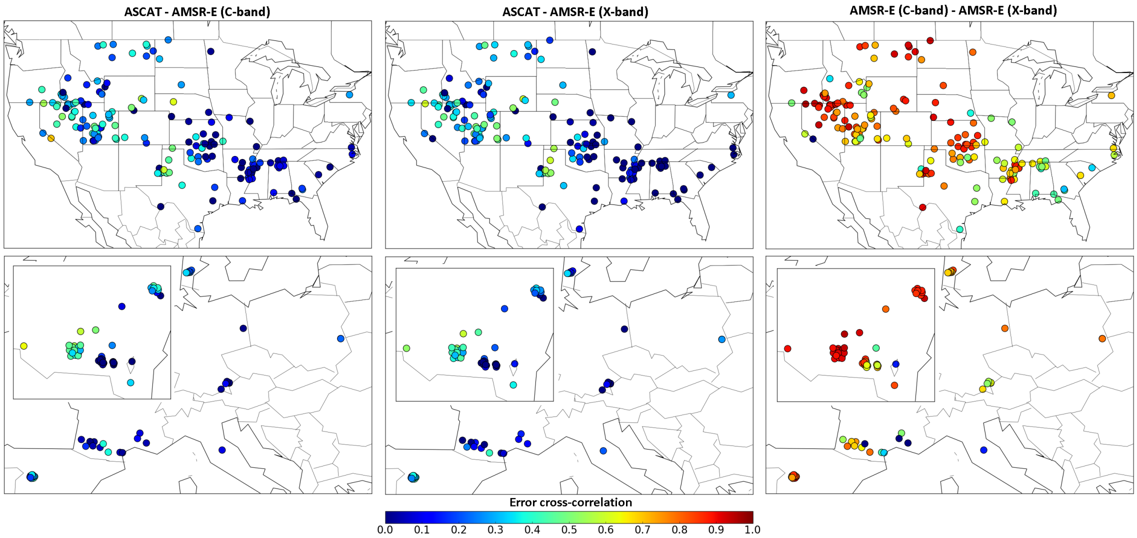

Figure 4.4: *Estimated cross-correlations between the errors in anomalies of ASCAT and AMSR-E C-band channel (left), ASCAT and AMSR-E X-band channel (middle), and both AMSR-E C- and X-band channels (right) for the Contiguous United States, Europe, and New South Wales (Australia).*

we investigate the impact of representativeness errors on error cross-correlation estimates in EC analysis analytically.

4.3.2.2 Representativeness errors in EC analysis

Following Gruber et al. (2015) we can split the observed soil moisture signal Θ into a joint signal component Θ_j, which is observed by all data sets, and a coarse-scale component Θ_c, which is observed by the coarse-resolution data sets only. Let us now consider four data sets a, b, c and d, where a represents a point-scale in situ data set, and the others represent data sets with comparable coarse spatial resolution - such as for instance GLDAS-Noah, ASCAT, and AMSR-E - with the errors between data sets c and d being correlated. The covariances between the data sets then write as:

$$
\begin{aligned}
\sigma_{ab} &= \beta_a \beta_b \sigma^2_{\Theta_j} \\
\sigma_{ac} &= \beta_a \beta_c \sigma^2_{\Theta_j} \\
\sigma_{ad} &= \beta_a \beta_d \sigma^2_{\Theta_j} \\
\sigma_{bc} &= \beta_b \beta_c (\sigma^2_{\Theta_j} + \sigma^2_{\Theta_c}) \\
\sigma_{bd} &= \beta_b \beta_d (\sigma^2_{\Theta_j} + \sigma^2_{\Theta_c}) \\
\sigma_{cd} &= \beta_c \beta_d (\sigma^2_{\Theta_j} + \sigma^2_{\Theta_c}) + \sigma_{\varepsilon_c \varepsilon_d}
\end{aligned}
\tag{4.1}
$$

From (4.1) we can see that the error cross-covariance estimators $\sigma_{\varepsilon_c\varepsilon_d} = \sigma_{cd} - \frac{\sigma_{ac}\sigma_{bd}}{\sigma_{ab}} = \sigma_{cd} - \frac{\sigma_{ad}\sigma_{bc}}{\sigma_{ab}}$ in (4.6) remain unbiased. That is, even though non-zero error cross-correlations between ASCAT and AMSR-E are observed mainly in areas where in situ stations are expected to have limited representativeness, these representativeness errors should not induce biases in error cross-correlation estimates. Instead, the same phenomena that decrease spatial representativeness of point measurements, i.e., highly localized soil moisture variations, might also induce correlations between retrieval errors of different satellites.

4.3.2.3 Global EC analysis

In Section 4.3.2.1, spatially limited in situ data were required to estimate error cross-correlations between the errors of ASCAT and AMSR-E. Here we will exclude the in situ data from EC analysis in order to estimate error cross-correlations between the AMSR-E products globally, yet it requires the assumption of zero error cross-correlation between ASCAT and AMSR-E. Even though we found that this assumption is not always fulfilled, observed non-zero error cross-correlations between ASCAT and AMSR-E are in general rather low (median ≈ 0.25) compared to those between the two AMSR-E products (median ≈ 0.8). Therefore - keeping a possible violation in mind - we will assume the cross-correlations between ASCAT and AMSR-E to be negligible.

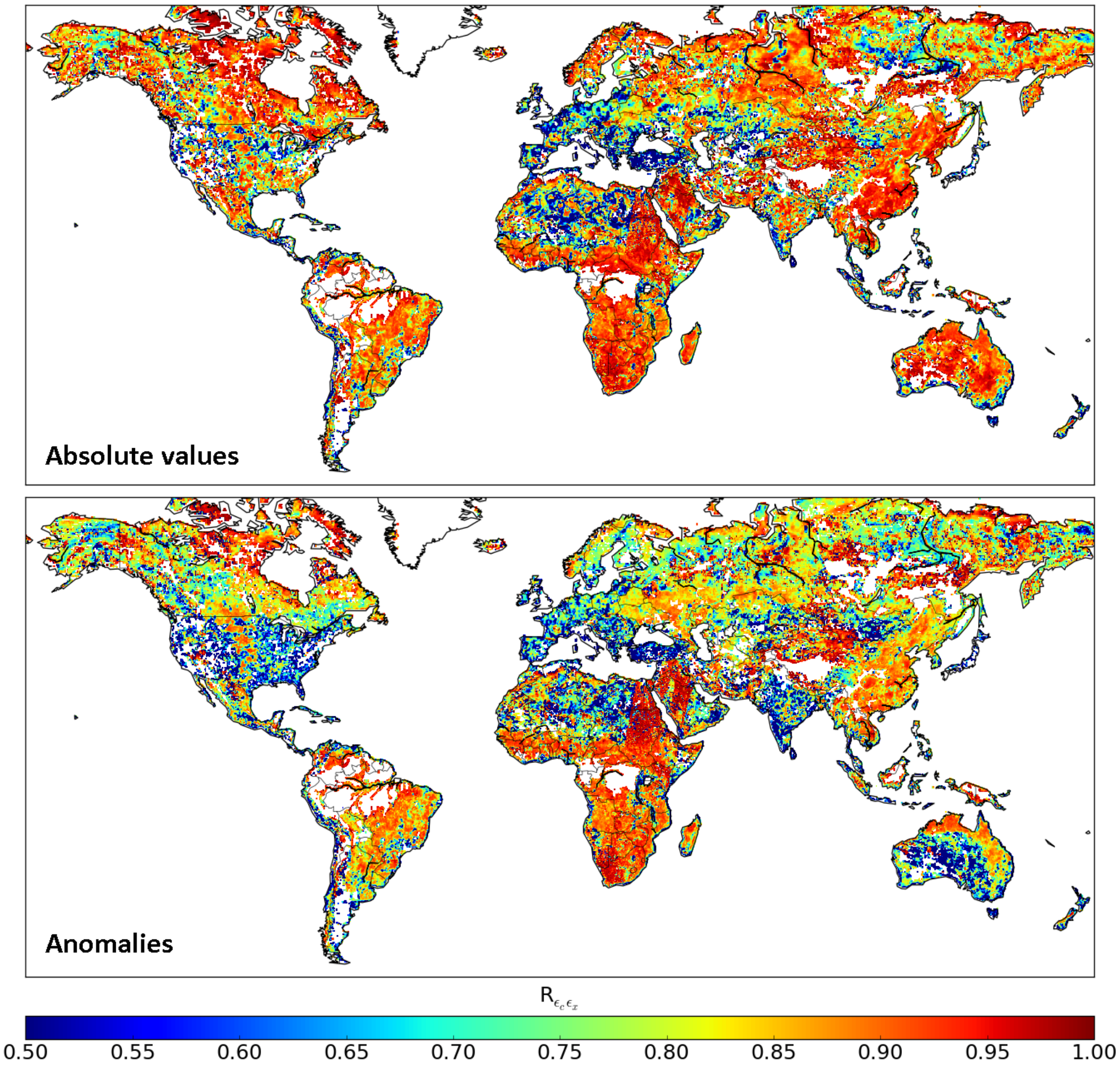

Figure 4.5: *Global error cross-correlation estimates for AMSR-E C- and X-band soil moisture retrievals. White shading indicates areas where estimates did not converge to a meaningful value.*

Figure 4.5 shows the error cross-correlation estimates between the C- and X-band soil moisture retrievals from AMSR-E for both absolute values and anomalies. White shading indicates areas where estimates did not converge to a meaningful value (i.e., where the cross-correlation estimate was below -1.0 or above 1.0 [-]). As already observed in the in situ analysis, very high error cross-correlations exist in most regions. The 5, 25, 50, 75, and 95% quantiles are 0.42, 0.76, 0.87, 0.92, and 0.97 [-] for absolute values, and 0.16, 0.70, 0.82, 0.90, and 0.99 [-] for anomalies, respectively. As mentioned before, these error cross-correlation estimates might be biased due to the presence of non-zero error cross-correlations between ASCAT and AMSR-E. However, the average value ranges are comparable to those obtained in Section 4.3.2.1, where globally-distributed in situ measurements were included as a fifth data set in EC analysis so that the error cross-correlation estimates for the AMSR-E products remain unaffected by non-zero error cross-correlations between ASCAT and AMSR-E. This suggests that the possible biases in the AMSR-E C- and X-band error cross-correlation estimates from the global EC analysis presented in this section are largely negligible.

Clear spatial patterns exist which suggest that the method is not overly sensitive to estimation noise, which is expected given the large number of temporally matching observations (median: 781). Likely drivers for these apparent error cross-correlation patterns are the differing spatial resolution and penetration depth of the two AMSR-E frequency channels, their differing sensitivity to vegetation, topographic complexity and possibly also other land cover features, and - most importantly - radio frequency interference (RFI). Indeed, regions with low error cross-correlation show good agreement with regions where RFI is expected (*de Nijs et al.*, 2015): C-band RFI contamination is expected mainly in the US, the Middle East and Japan, whereas X-band RFI is expected mainly over England and Italy. RFI in both frequencies is also expected in Europe, especially around densely urbanized areas. In most of these regions, also lower error cross-correlations are observed. This good agreement is a first indicator for the reliability of EC error cross-correlation estimates. However, additional validation is required before the approach can be applied with full confidence.

4.4 Summary and outlook

A method for estimating error cross-correlations between soil moisture data sets was developed by generalizing the well-known triple collocation (TC) analysis to an arbitrary number of data sets and relaxing the assumption of non-zero error cross-correlation for some data set combinations, referred to as extended collocation (EC) analysis. The number of allowed non-zero error cross-correlations between data set pairs is mainly limited by the overall number of data sets used and by their underlying error cross-correlation structure: Each member of the data set pairs with assumed non-zero error cross correlation must also be a member of at least one data set triplet with fully independent errors. Furthermore, remaining degrees of freedom can be used to solve the collocation system of equations in a least-squares sense.

The proposed EC method was evaluated using both a synthetic identical twin experiment and real data experiments. In the synthetic experiment, EC analysis was able to recover true error cross-correlation levels with an average RMSD of 0.08 [-] and a negligible bias. In the real data experiments EC analysis was applied to satellite-based soil moisture retrievals from ASCAT, the AMSR-E C-band channel, the AMSR-E X-band channel, modelled soil moisture estimates from GLDAS-Noah, and in situ soil moisture measurements drawn from the International Soil Moisture Network. Results suggest that significant error cross-correlations exist between the AMSR-E C-band and X-band channels (median = 0.82 and 0.78 [-] for absolute values and anomalies, respectively), which are likely driven by their differing spatial resolution, sampling depth, sensitivity to vegetation and other land cover features, and - most importantly - RFI. Moreover, slight non-zero error cross-correlations were found also between ASCAT and AMSR-E (median = 0.25 and 0.20 [-] for absolute values and anomalies, respectively). These non-zero error cross-correlations may slightly bias the error cross-correlation estimates between the AMSR-E C- and X-band channels.

It should be emphasized that - even though only demonstrated for four and five data sets - the EC method presented in this study is readily applicable to an arbitrary number of data sets, which would facilitate the estimation of more non-zero error cross-covariance terms (e.g., when using 3 passive data sets such as SMAP, AMSR2, and SMOS together with 2 active data sets such as MetOp-A and MetOp-B). Therefore, it represents an important step towards a fully-parameterized error covariance matrix which is vital for any rigorous data assimilation framework or data merging scheme.

Bibliography

Albergel, C., C. Ruediger, T. Pellarin, J. Calvet, N. Fritz, F. Froissard, D. Suquia, A. Petitpa, B. Piguet, and E. Martin (2008), From near-surface to root-zone soil moisture using an exponential filter: an assessment of the method based on in-situ observations and model simulations., *Hydrology and earth system sciences.*, **12**(6), p. 1323–1337. 74

Bell, J. E., M. A. Palecki, C. B. Baker, W. G. Collins, J. H. Lawrimore, R. D. Leeper, M. E. Hall, J. Kochendorfer, T. P. Meyers, T. Wilson, et al. (2013), Us climate reference network soil moisture and temperature observations, *Journal of Hydrometeorology*, **14**(3), p. 977–988. 74

Bircher, S., N. Skou, K. H. Jensen, J. Walker, and L. Rasmussen (2012), A soil moisture and temperature network for smos validation in western denmark, *Hydrology and Earth System Sciences*, **16**(5), p. 1445–1463. 74

Bolten, J., and W. Crow (2012), Improved prediction of quasi-global vegetation conditions using remotely-sensed surface soil moisture, *Geophysical Research Letters*, **39**(19). 42, 66

Bowden, R. J., and D. A. Turkington (1990), *Instrumental variables*, vol. 8, Cambridge University Press. 14

Brocca, L., S. Hasenauer, T. Lacava, F. Melone, T. Moramarco, W. Wagner, W. Dorigo, P. Matgen, J. Martinez-Fernandez, P. Llorens, J. Latron, C. Martin, and M. Bittelli (2011), Soil moisture estimation through ascat and amsr-e sensors: An intercomparison and validation study across europe, *Remote Sensing of Environment*, **115**(12), p. 3390–3408. 74

Brocca, L., T. Moramarco, F. Melone, W. Wagner, S. Hasenauer, and S. Hahn (2012), Assimilation of surface-and root-zone ascat soil moisture products into rainfall–runoff modeling, *Geoscience and Remote Sensing, IEEE Transactions on*, **50**(7), p. 2542–2555. 42

Caires, S., and A. Sterl (2003), Validation of ocean wind and wave data using triple collocation, *Journal of Geophysical Research: Oceans (1978–2012)*, **108**(C3). 4, 12, 67

Crow, W., and M. Van den Berg (2010), An improved approach for estimating observation and model error parameters in soil moisture data assimilation, *Water Resources Research*, **46**(12). 4, 20, 42, 66

Crow, W., and M. T. Yilmaz (2014), The auto-tuned land data assimilation system (atlas), *Water Resources Research*, **50**(1), p. 371–385. 42, 45, 66

Crow, W., S. Kumar, and J. Bolten (2012a), On the utility of land surface models for agricultural drought monitoring, *Hydrology and Earth System Sciences*, **16**(9), p. 3451–3460. 43, 72

Crow, W. T., and E. Van Loon (2006), Impact of incorrect model error assumptions on the sequential assimilation of remotely sensed surface soil moisture, *Journal of hydrometeorology*, **7**(3), p. 421–432. 42

Crow, W. T., and E. F. Wood (2002), Impact of soil moisture aggregation on surface energy flux prediction during sgp'97, *Geophysical research letters*, **29**(1), p. 8–1. 18

Crow, W. T., R. D. Koster, R. H. Reichle, and H. O. Sharif (2005), Relevance of time-varying and time-invariant retrieval error sources on the utility of spaceborne soil moisture products, *Geophysical Research Letters*, **32**(24). 19

Crow, W. T., A. A. Berg, M. H. Cosh, A. Loew, B. P. Mohanty, R. Panciera, P. de Rosnay, D. Ryu, and J. P. Walker (2012b), Upscaling sparse ground-based soil moisture observations for the validation of coarse-resolution satellite soil moisture products, *Rev. Geophys.*, **50**(2), p. RG2002. 18, 21, 75

Crow, W. T., C.-H. Su, D. Ryu, and M. T. Yilmaz (2015), Optimal averaging of soil moisture predictions from ensemble land surface model simulations, *Water Resources Research*, **51**(11), p. 9273–9289. 4, 66

De Lannoy, G. J., P. R. Houser, N. E. Verhoest, V. R. Pauwels, and T. J. Gish (2007), Upscaling of point soil moisture measurements to field averages at the ope 3 test site, *Journal of Hydrology*, **343**(1), p. 1–11. 17

de Nijs, A. H., R. M. Parinussa, R. A. de Jeu, J. Schellekens, and T. R. Holmes (2015), A methodology to determine radio-frequency interference in amsr2 observations, *Geoscience and Remote Sensing, IEEE Transactions on*, **53**(9), p. 5148–5159. 79

de Rosnay, P., M. Drusch, D. Vasiljevic, G. Balsamo, C. Albergel, and L. Isaksen (2013), A simplified extended kalman filter for the global operational soil moisture analysis at ecmwf, *Quarterly Journal of the Royal Meteorological Society*, **139**(674), p. 1199–1213. 42, 66

Dorigo, W., P. van Oevelen, W. Wagner, M. Drusch, S. Mecklenburg, A. Robock, and T. Jackson (2011a), A new international network for in situ soil moisture data, *Eos Transactions AGU*, **92**(17), p. 141–142. 3, 74

Dorigo, W., A. Xaver, M. Vreugdenhil, A. Gruber, H. A, A. Sanchis-Dufau, D. Zamojski, C. Cordes, W. Wagner, and M. Drusch (2013), Global automated quality control of in situ soil moisture data from the international soil moisture network, *Vadose Zone Journal*, **12**(3). 52, 74

Dorigo, W. A., K. Scipal, R. M. Parinussa, Y. Y. Liu, W. Wagner, R. A. M. de Jeu, and V. Naeimi (2010), Error characterisation of global active and passive microwave soil moisture datasets, *Hydrol. Earth Syst. Sci.*, **14**(12), p. 2605–2616. 4, 12, 14, 20, 42, 68

Dorigo, W. A., W. Wagner, R. Hohensinn, S. Hahn, C. Paulik, A. Xaver, A. Gruber, M. Drusch, S. Mecklenburg, P. van Oevelen, A. Robock, and T. Jackson (2011b), The international soil moisture network: a data hosting facility for global in situ soil moisture measurements, *Hydrol. Earth Syst. Sci.*, **15**(5), p. 1675–1698. 52, 74

Draper, C., R. Reichle, R. de Jeu, V. Naeimi, R. Parinussa, and W. Wagner (2013), Estimating root mean square errors in remotely sensed soil moisture over continental scale domains, *Remote Sensing of Environment*, **137**, p. 288–298. 13, 18, 20, 24, 28

Drusch, M., E. Wood, and H. Gao (2005), Observation operators for the direct assimilation of trmm microwave imager retrieved soil moisture, *Geophysical Research Letters*, **32**(15). 17, 18

Entekhabi, D., R. H. Reichle, R. D. Koster, and W. T. Crow (2010), Performance metrics for soil moisture retrievals and application requirements, *J. Hydrometeor*, **11**(3), p. 832–840, doi:10.1175/2010JHM1223.1. 22

Fang, H., S. Wei, C. Jiang, and K. Scipal (2012), Theoretical uncertainty analysis of global modis, cyclopes, and globcarbon lai products using a triple collocation method, *Remote Sensing of Environment*, **124**, p. 610–621. 4, 12, 67

Gruber, A., W. Dorigo, S. Zwieback, A. Xaver, and W. Wagner (2013), Characterizing coarse-scale representativeness of in situ soil moisture measurements from the international soil moisture network, *Vadose Zone Journal*, **12**(2). 3, 21, 23, 75

Gruber, A., W. Crow, W. Dorigo, and W. Wagner (2015), The potential of 2d kalman filtering for soil moisture data assimilation, *Remote Sensing of Environment*, **171**, p. 137–148. 6, 68, 72, 75, 77

Gruber, A., C.-H. Su, S. Zwieback, W. Crow, W. Dorigo, and W. Wagner (2016a), Recent advances in (soil moisture) triple collocation analysis, *International Journal of Applied Earth Observation and Geoinformation*, **45**, p. 200–211. 4, 6

Gruber, A., C.-H. Su, W. Crow, S. Zwieback, W. Dorigo, and W. Wagner (2016b), Estimating error cross-correlations in soil moisture data sets using extended collocation analysis, *Journal of Geophysical Research: Atmospheres*, **121(3)**, p. 1208–1219. 6, 7

Gupta, H. V., H. Kling, K. K. Yilmaz, and G. F. Martinez (2009), Decomposition of the mean squared error and nse performance criteria: Implications for improving hydrological modelling, *Journal of Hydrology*, **377**(1), p. 80–91. 23, 36

Han, X., X. Li, H. Hendricks Franssen, H. Vereecken, and C. Montzka (2012), Spatial horizontal correlation characteristics in the land data assimilation of soil moisture, *Hydrology and Earth System Sciences*, **16**(5), p. 1349–1363. 62

Hollinger, S. E., and S. A. Isard (1994), A soil moisture climatology of illinois, *Journal of Climate*, **7**(5), p. 822–833. 74

Huffman, G. J., R. F. Adler, D. T. Bolvin, and E. J. Nelkin (2010), The trmm multi-satellite precipitation analysis (tmpa), in *Satellite rainfall applications for surface hydrology*, p. 3–22, Springer. 51

Jackson, T., M. Cosh, R. Bindlish, P. Starks, D. Bosch, M. Seyfried, D. Goodrich, M. Moran, and J. Du (2010), Validation of advanced microwave scanning radiometer soil moisture products, *Geoscience and Remote Sensing, IEEE Transactions on*, **48**(12), p. 4256–4272. 3, 23, 52, 74

Kerr, Y., P. Waldteufel, J.-P. Wigneron, S. Delwart, F. Cabot, J. Boutin, M. Escorihuela, J. Font, N. Reul, C. Gruhier, S. Juglea, M. Drinkwater, A. Hahne, M. Martin-Neira, and S. Mecklenburg (2010), The smos mission: New tool for monitoring key elements ofthe global water cycle, *Proceedings of the IEEE*, **98**(5), p. 666–687. 42

Larson, K. M., E. E. Small, E. D. Gutmann, A. L. Bilich, J. J. Braun, and V. U. Zavorotny (2008), Use of gps receivers as a soil moisture network for water cycle studies, *Geophysical Research Letters*, **35**(24). 74

Leavesley, G., O. David, D. Garen, J. Lea, J. Marron, T. Pagano, T. Perkins, and M. Strobel (2008), A modeling framework for improved agricultural water supply forecasting, in *AGU Fall Meeting Abstracts*, vol. 1, p. 0497. 74

Legates, D. R., R. Mahmood, D. F. Levia, T. L. DeLiberty, S. M. Quiring, C. Houser, and F. E. Nelson (2011), Soil moisture: A central and unifying theme in physical geography, *Progress in Physical Geography*, **35**(1), p. 65–86, doi:10.1177/0309133310386514. 3, 12, 66

Liu, Q., R. H. Reichle, R. Bindlish, M. H. Cosh, W. T. Crow, R. de Jeu, G. J. De Lannoy, G. J. Huffman, and T. J. Jackson (2011a), The contributions of precipitation and soil moisture observations to the skill of soil moisture estimates in a land data assimilation system, *Journal of Hydrometeorology*, **12**(5), p. 750–765. 57

Liu, Y., W. Dorigo, R. Parinussa, R. de Jeu, W. Wagner, M. McCabe, J. Evans, and A. van Dijk (2012), Trend-preserving blending of passive and active microwave soil moisture retrievals, *Remote Sensing of Environment*, **123**(0), p. 280–297, doi:10.1016/j.rse.2012.03.014. 66

Liu, Y. Y., R. M. Parinussa, W. A. Dorigo, R. A. M. De Jeu, W. Wagner, A. I. J. M. van Dijk, M. F. McCabe, and J. P. Evans (2011b), Developing an improved soil moisture dataset by blending passive and active microwave satellite-based retrievals, *Hydrol. Earth Syst. Sci.*, **15**(2), p. 425–436. 3, 12, 66

Loew, A., and F. Schlenz (2011), A dynamic approach for evaluating coarse scale satellite soil moisture products, *Hydrol. Earth Syst. Sci.*, **15**(1), p. 75–90. 12, 14, 15, 19

Marczewski, W., J. Slominski, E. Slominska, B. Usowicz, J. Usowicz, S. Romanov, O. Maryskevych, J. Nastula, and J. Zawadzki (2010), Strategies for validating and directions for employing smos data, in the cal-val project swex (3275) for wetlands, *Hydrol. Earth Syst. Sci. Discuss.*, **7**(5), p. 7007–7057. 74

McColl, K. A., J. Vogelzang, A. G. Konings, D. Entekhabi, M. Piles, and A. Stoffelen (2014), Extended triple collocation: Estimating errors and correlation coefficients with respect to an unknown target, *Geophysical Research Letters*, **41**(17), p. 6229–6236. 12, 13, 14, 15, 25, 45

Miralles, D. G., W. T. Crow, and M. H. Cosh (2010), Estimating spatial sampling errors in coarse-scale soil moisture estimates derived from point-scale observations, *J. Hydrometeor*, **11**(6), p. 1423–1429, doi:10.1175/2010JHM1285.1. 3, 18, 21, 75

Naeimi, K. B. Z. H. S. . W. W., Vahid; Scipal (2009), An improved soil moisture retrieval algorithm for ers and metop scatterometer observations, *Geoscience and Remote Sensing, IEEE Transactions on*, **47**(7), p. 1999–2013. 27, 42, 73

Naeimi, V., C. Paulik, A. Bartsch, W. Wagner, R. Kidd, S.-E. Park, K. Elger, and J. Boike (2012), Ascat surface state flag (ssf): Extracting information on surface freeze/thaw conditions from backscatter data using an empirical threshold-analysis algorithm, *Geoscience and Remote Sensing, IEEE Transactions on*, **50**(7), p. 2566–2582. 27, 51, 74

Owe, M., R. de Jeu, and T. Holmes (2008), Multisensor historical climatology of satellite-derived global land surface moisture, *Journal of Geophysical Research: Earth Surface*, **113**. 27, 73

Pan, M., C. K. Fisher, N. W. Chaney, W. Zhan, W. T. Crow, F. Aires, D. Entekhabi, and E. F. Wood (2015), Triple collocation: Beyond three estimates and separation of structural/non-structural errors, *Remote Sensing of Environment*, **171**, p. 299–310. 66

Parinussa, R. M., A. G. Meesters, Y. Y. Liu, W. Dorigo, W. Wagner, and R. A. De Jeu (2011), Error estimates for near-real-time satellite soil moisture as derived from the land parameter retrieval model, *Geoscience and Remote Sensing Letters, IEEE*, **8**(4), p. 779–783. 73

Pellarin, T., J.-P. Laurent, B. Cappelaere, B. Decharme, L. Descroix, and D. Ramier (2009), Hydrological modelling and associated microwave emission of a semi-arid region in south-western niger, *Journal of Hydrology*, **375**(1), p. 262–272. 74

Pierdicca, N., F. Fascetti, L. Pulvirenti, R. Crapolicchio, and J. Munoz-Sabater (2015), Quadruple collocation analysis for soil moisture product assessment, *Geoscience and Remote Sensing Letters, IEEE*, **12**(8), p. 1595–1599. 4, 67, 69

Reichle, R. H., and R. D. Koster (2003), Assessing the impact of horizontal error correlations in background fields on soil moisture estimation, *Journal of Hydrometeorology*, **4**(6), p. 1229–1242. 42, 62

Reichle, R. H., D. B. McLaughlin, and D. Entekhabi (2002), Hydrologic data assimilation with the ensemble kalman filter, *Monthly Weather Review*, **130**(1), p. 103–114. 12

Reichle, R. H., W. T. Crow, and C. L. Keppenne (2008), An adaptive ensemble kalman filter for soil moisture data assimilation, *Water resources research*, **44**(3). 42

Rodell, M., P. Houser, U. e. a. Jambor, J. Gottschalck, K. Mitchell, C. Meng, K. Arsenault, B. Cosgrove, J. Radakovich, M. Bosilovich, et al. (2004), The global land data assimilation system, *Bulletin of the American Meteorological Society*, **85**(3), p. 381–394. 74

Roebeling, R., E. Wolters, J. Meirink, and H. Leijnse (2012), Triple collocation of summer precipitation retrievals from seviri over europe with gridded rain gauge and weather radar data, *Journal of Hydrometeorology*, **13**(5), p. 1552–1566. 4, 12, 67

Schlenz, F., J. T. Dall'Amico, A. Loew, and W. Mauser (2012), Uncertainty assessment of the smos validation in the upper danube catchment, *Geoscience and Remote Sensing, IEEE Transactions on*, **50**(5), p. 1517–1529. 74

Scipal, K., T. Holmes, R. de Jeu, V. Naeimi, and W. Wagner (2008), A possible solution for the problem of estimating the error structure of global soil moisture data sets, *Geophys. Res. Lett.*, **35**(24), p. L24,403. 4, 12, 14, 20, 42, 68

Smith, A., J. Walker, A. Western, R. Young, K. Ellett, R. Pipunic, R. Grayson, L. Siriwardena, F. Chiew, and H. Richter (2012), The murrumbidgee soil moisture monitoring network data set, *Water Resources Research*, **48**(7). 74

Stoffelen, A. (1998), Toward the true near-surface wind speed: Error modeling and calibration using triple collocation, *J. Geophys. Res.*, **103**(C4), p. 7755–7766. 4, 12, 13, 14, 15, 17, 18, 36, 37, 42, 44, 45, 66, 68

Su, C.-H., and D. Ryu (2015), Multi-scale analysis of bias correction of soil moisture, *Hydrology and Earth System Sciences*, **19**(1), p. 17–31. 17, 18

Su, C.-H., D. Ryu, W. T. Crow, and A. W. Western (2014a), Beyond triple collocation: Applications to soil moisture monitoring, *Journal of Geophysical Research: Atmospheres*, **119**(11), p. 6419–6439. 4, 14, 15, 19, 66, 69

Su, C.-H., D. Ryu, W. T. Crow, and A. W. Western (2014b), Stand-alone error characterisation of microwave satellite soil moisture using a fourier method, *Remote Sensing of Environment*, **154**, p. 115–126. 12, 13, 14, 15, 25, 68

Su, Z., J. Wen, L. Dente, R. van der Velde, L. Wang, Y. Ma, K. Yang, and Z. Hu (2011), The tibetan plateau observatory of plateau scale soil moisture and soil temperature (tibet-obs) for quantifying uncertainties in coarse resolution satellite and model products, *Hydrol. Earth Syst. Sci.*, **15**(7), p. 2303–2316. 74

Vachaud, G., A. Passerat De Silans, P. Balabanis, and M. Vauclin (1985), Temporal stability of spatially measured soil water probability density function, *Soil Sci. Soc. Am. J.*, **49**(4), p. 822–828, doi:10.2136/sssaj1985.03615995004900040006x. 3, 21

Vogelzang, J., and A. Stoffelen (2012), Triple collocation, *EUMETSAT Report.[Available at http://research.metoffice.gov.uk/research/interproj/nwpsaf/scatterometer/TripleCollocation_NWPSAF_TR_KN_021_v1_0.pdf.]*. 37, 75

Vogelzang, J., A. Stoffelen, A. Verhoef, and J. Figa-Saldaña (2011), On the quality of high-resolution scatterometer winds, *Journal of Geophysical Research: Oceans (1978–2012)*, **116**(C10). 4, 12, 67

Wagner, W., G. Lemoine, and H. Rott (1999), A method for estimating soil moisture from ers scatterometer and soil data, *Remote Sensing of Environment*, **70**(2), p. 191–207, doi:10.1016/S0034-4257(99)00036-X. 27, 42, 73

Wagner, W., C. Pathe, M. Doubkova, D. Sabel, A. Bartsch, S. Hasenauer, G. Blöschl, K. Scipal, J. Martínez-Fernández, and A. Löw (2008), Temporal stability of soil moisture and radar backscatter observed by the advanced synthetic aperture radar (asar), *Sensors*, **8**(2), p. 1174–1197. 21

Yilmaz, M., W. Crow, M. Anderson, and C. Hain (2012), An objective methodology for merging satellite-and model-based soil moisture products, *Water Resources Research*, **48**(11). 66

Yilmaz, M. T., and W. T. Crow (2013), The optimality of potential rescaling approaches in land data assimilation., *Journal of Hydrometeorology*, **14**(2). 15, 17

Yilmaz, M. T., and W. T. Crow (2014), Evaluation of assumptions in soil moisture triple collocation analysis, *Journal of Hydrometeorology*, **15**(3), p. 1293–1302. 4, 20, 69

Young, R., J. Walker, N. Yeoh, A. Smith, K. Ellett, O. Merlin, and A. Western (2008), Soil moisture and meteorological observations from the murrumbidgee catchment, *Department of Civil and Environmental Engineering, The University of Melbourne*. 74

Zreda, M., D. Desilets, T. Ferré, and R. L. Scott (2008), Measuring soil moisture content non-invasively at intermediate spatial scale using cosmic-ray neutrons, *Geophysical Research Letters*, **35**(21). 74

Zwieback, S., K. Scipal, W. Dorigo, and W. Wagner (2012), Structural and statistical properties of the collocation technique for error characterization, *Nonlin. Processes Geophys.*, **19**(1), p. 69–80. 4, 19, 46, 67, 69

Zwieback, S., W. Dorigo, and W. Wagner (2013), Estimation of the temporal autocorrelation structure by the collocation technique with an emphasis on soil moisture studies, *Hydrological Sciences Journal*, **58**(8), p. 1729–1747. 4, 19

PERSONAL INFORMATION

Dipl.-Ing. Alexander Gruber

📍 Grenzgasse 14-18/42, 2340 Moedling, Austria

☎ +43 1 58801 12263 📱 +43 699 1273 9053

✉ alexander.gruber@geo.tuwien.ac.at

🌐 http://geo.tuwien.ac.at/staff/alexander-gruber/

Sex Male | Date of birth 28/02/1988 | Nationality Austria

POSITION

Researcher

WORK EXPERIENCE

12/2009 – present

Project Assistant/University Assistant

Technische Universitaet Wien (TU Wien), Department of Geodesy and Geoinformation (GEO)
Vienna, Austria

- Research in the field of microwave remote sensing
- Software development
- Project coordination
- Teaching

Business or sector Higher Education and Research

06/2015 – 09/2015

Research Assistent

University of Melbourne, Department of Infrastructure Engineering
Melbourne, Victoria, Australia

- Research in the field of microwave remote sensing
- Teaching

Business or sector Higher Education and Research

07/2014 – 10/2014

Visiting Scientist

United States Department of Agriculture (USDA), Hydrology and Remote Sensing Laboratory
Beltsville, Maryland, USA

- Research in the field of microwave remote sensing

Business or sector Research

EDUCATION AND TRAINING

11/2013 – 04/2016

Doctoral program in Engineering Sciences
Dissertation field: Surveying and Geoinformation

EQF Level 8

Technische Universitaet Wien, Vienna, Austria

- Microwave Remote Sensing
- Research methodology
- Scientific writing
- Software development

04/2012 – 11/2013

Master of Science in Geodesy and Geophysics

EQF Level 7

Technische Universitaet Wien, Vienna, Austria

- Microwave Remote Sensing
- Statistics
- Software development
- Basic knowledge in research methodology

© European Union, 2002-2015 | europass.cedefop.europa.eu

10/2008 – 04/2012

Bachelor of Science in Geodesy and Geoinformation

EQF Level 6

Technische Universitaet Wien, Vienna, Austria

- Physics
- Mathematics
- Software development
- Basic knowledge in remote sensing

PERSONAL SKILLS

Mother tongue(s)

German

Other language(s)

	UNDERSTANDING		SPEAKING		WRITING
	Listening	Reading	Spoken interaction	Spoken production	
English	C2	C2	C2	C2	C2

Communication skills

- Good communication skills gained by working in several international project teams.
- Good presentation skills gained from many oral presentations at international conferences and project meetings

Organisational / managerial skills

- Project management experience as technical and scientific coordinator of several research and development projects
- Experience in the preparation of research proposals.

Technical Skills

- Enhanced programming skills (Python, IDL, MATLAB, R, Assembler, C/C++)
- Remote sensing software (ENVI, NEST, OPALS)
- GIS software (ArcGIS, QGIS)
- Web development (PHP, PostgreSQL, HTML, JavaScript, Joomla!)
- Operating systems (Windows, Linux, Mac OS)
- Word processing software (LaTeX, MS Office)

ADDITIONAL INFORMATION

Publications

- Authored and co-authored many peer-reviewed journal papers, book chapters, and conference proceedings.

- Selected publications:

    - **Gruber, A**., Dorigo, W. A., Zwieback, S., Xaver, A., & Wagner, W. (2013). Characterizing coarse-scale representativeness of in situ soil moisture measurements from the International Soil Moisture Network. *Vadose Zone Journal*, *12*(2).

    - **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*, 45, 200-211.

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

    - **Gruber, A**., Crow, W., Dorigo, W., & Wagner, W. (2015). The potential of 2D Kalman filtering for soil moisture data assimilation. *Remote Sensing of Environment*, 171, 137-148.

- Full list of publications: https://scholar.google.at/citations?user=s86d3XMAAAAJ&hl=de

Projects

- Experience as technical coordinator, scientific coordinator or key science participant in several international projects.

- Selected projects:

  - **SCA cross-polarisation**
    Supporting agency: EUMETSAT
    Duties: Investigating the feasibility of using cross-polarized satellite radar signals for improving soil moisture retrievals through an improved vegetation characterization.

  - **H-SAF Second Continuous Development and Operations Phase (CDOP-2)**
    Supporting agency: EUMETSAT
    Duties: Processing and improvement of Synthetic Aperture Radar (SAR) data. Satellite product validation studies.

  - **Climate Change Initiative (CCI) for Soil Moisture**
    Supporting agency: ESA
    Duties: Quality inter-comparison of various soil moisture retrieval algorithms for active and passive microwave sensors to assess their suitability for being integrated into a blended product. Development and application of an adaptive filtering approach to characterize errors in the Essential Climate Variable (ECV) soil moisture data set that is being created within the CCI.

  - **eartH2Observe**
    Supporting agency: European Commission (FP7)
    Duties: Development of novel error characterization methods for satellite-based soil moisture data sets with a focus on harmonization and blending of multiple data sources to a consistent long-term soil moisture record.

© European Union, 2002-2015 | europass.cedefop.europa.eu