TU Wien
Department of Geodesy and Geoinformation
Gußhausstraße 27-29/E120
1040 Vienna
Austria

Diplomarbeit

Evaluation and Enhancement of Automated Quality Control Procedures

for the International Soil Moisture Network

Elsa Heer

2017

betreut durch

Projektass. Dipl.-Ing. Angelika Xaver

unter der Anleitung von

Senior Scientist Dr.rer.nat. Wouter Arnoud Dorigo MSc

Abstract

The quality of in situ soil moisture data is of high importance since it is still the most trusted source for satellite data validations. An erroneous behavior of in situ soil moisture data is very difficult to detect, due to annual and daily changes and most significantly the high influence of precipitation and snow melting processes. The International Soil Moisture Network (ISMN) provides in situ soil moisture data sets from all around the world, from different data providers, observed with different sensors in different depths. The data processing routines of the ISMN already contain very sophisticated algorithms for the detection of erroneous data. Since the development of these algorithms, many more data sets were added to the ISMN and new types of erroneous observations could be identified. Thus, a revision and extension of the existing algorithms became necessary. For this thesis the algorithms for the error detection were adapted and additionally, new methods of error detection were developed. To evaluate the revised automated quality control system many in situ soil moisture timeseries were chosen and manually validated to be compared to the existing quality control procedures and the new algorithms. Improvements of the new algorithms will be shown to provide a valuable quality assessment of the ISMN data sets, which are the foundation of many scientific publications.

Kurzfassung

Die Qualitätskontrolle von ISMN Bodenfeuchtedaten ist von hoher Bedeutung, da sie noch immer als hoch vertrauenswürdige Basis für Validierungen von Satellitendaten gelten. Fehlerhafte Bodenfeuchtedaten sind äußerst komplex zu detektieren, da sie eine hohe Variation aufgrund von Tages und Jahreszyklen, aber vor allem durch Regenfälle und Schneeschmelze, aufweisen. Das International Soil Moisture Network (ISMN) bietet weltweite in situ Bodenfeuchte Daten von verschiedenen Datenanbietern, die mit verschiedenen Sensoren, in diversen Tiefen beobachtet wurden. Die Routinen zur Harmonisierung der verschiedenen Datensätze des ISMN beinhalten bereits höchst durchdachte Algorithmen zur Detektierung verdächtiger Daten. Seit der Entwicklung dieser Algorithmen wurde die Datenbank des ISMN um sehr viele Datensätze erweitert und zusätzliche Arten von Fehlern wurden identifiziert. Deshalb wurde eine Überarbeitung der existierenden Qualitätskontrolle notwendig. Im Rahmen dieser Arbeit wurden die Algorithmen zur Fehlerdetektion adaptiert und neue Methoden der Fehlererkennung entwickelt. Für die Evaluierung der überarbeiteten automatisierten Fehlerkontrolle wurden viele Datensätze gewählt und händisch validiert, um diese mit den Ergebnissen der existierenden und neuen Algorithmen zu vergleichen. Verbesserungen durch die adaptierten und neuen Algorithmen für eine wertvolle Qualitätsprüfung der ISMN Datensätze, die Grundlage vieler wissenschaftlicher Arbeiten sind, werden zu sehen sein.

List of Tables

List of Figures

Contents

Acknowledgements

First of all I thank Angelika for her previous work that I was allowed to continue and for her help during the preparation of this thesis. I also thank my supervisor Wouter, and my colleague Romina for their notion.

Most of all I thank my family, particularly my parents, for their unconditional support and patience during my studies.

Crucial for this thesis was the contribution of five International Soil Moisture Network (ISMN) data providers: Json Downing and Roger Ruess (BNZ-LTER), Udo Rummel (MOL-RAO), Heye Bogena (TERENO), Mahta Moghaddam, Daniel Clewley, Agnelo Silva and Ruzbeh Akba (SOILSCAPE) and Michael Palecki (USCRN), who provide their own quality assessment for their in situ data observations.

Last but not least I take this opportunity to acknowledge the valuable contribution of all the other ISMN data providers.

1 Introduction

This thesis is based on the existing work of Angelika Xaver [8], who developed a very innovative automated quality control system for ground based (in situ) soil moisture data provided by the International Soil Moisture Network (ISMN) [2] [3]. The ISMN was initiated by the European Space Agency (ESA) in 2010 to share soil moisture data from different data providers all over the world in one consistent format, retrievable in one data portal.

Erroneous behavior of in situ soil moisture data sets are very difficult to detect, due to annual and daily variations and most significantly the high influence of precipitation events and snow melting processes. Only few of the ISMN data providers have their own quality control system. These Quality assessments from different providers are also based on different methods and lead to inhomogeneous notations, which are complex to adapt for the user. Therefore, advanced quality control procedures were developed for the whole ISMN. A quality control system for data sets from places all around the world, with different land cover and climate classes, observed with different soil moisture sensors is especially complex, because of different soil moisture characteristics and therefore a large variety of data errors. Since the development of the automated quality control system, the ISMN became very popular and evolved very quickly. More and more people are willing to share their in situ soil moisture data with the ISMN and thus, making it available for a wide public and scientific studies. Therefore, the reliability of the ISMN quality assessment can now be tested on many more various data sets than at the time of their development. In addition, the available information of the quality assessment of the data providers will be analyzed, to find advantages that can be brought to use for the ISMN.

The main content of this thesis is parted in six chapters. The chapter 'Data' includes a description of in situ soil moisture behavior and measure methods, followed by an introduction of the ISMN. The chapter 'Existing Methods' consists of a detailed functional description of the current ISMN algorithms used for error detection. Some of the data providers of the ISMN also have their own quality control systems, which are integrated in the ISMN as so-called original quality flags. Those will also be discussed in the 'Existing Methods' chapter. In chapter 'New Methods' the new quality control methods will be introduced. In the chapter 'Evaluation' the performance of the new methods will be evaluated and compared to the existing algorithms. The last chapter 'Conclusion' will recap how the error detection could be enhanced with the revised algorithms and which further improvements might still be possible.

2 Data

2.1 Soil Moisture Dataset Characteristics

A soil moisture timeseries has a very typical spectrum. Since it is mostly affected by precipitation events it can rise very quickly within one hour, but the drying process may continue over a much longer time period of multiple hours or even days (Figure 2.1).

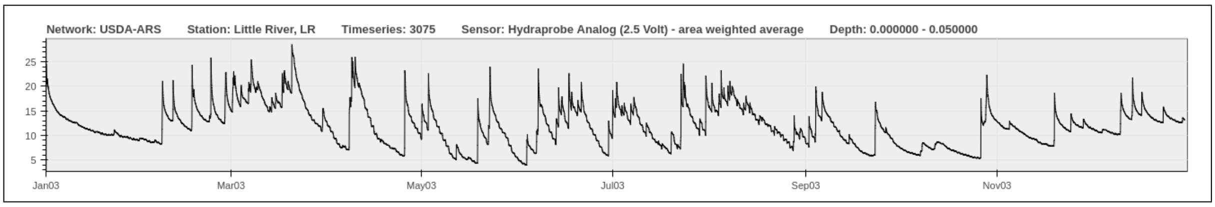


Figure 2.1: typical soil moisture data set without errors

During summer daily variations are often higher due to a higher variation of air temperature. When the soil is frozen in winter the data series shows a more constant signal.

Typical algorithms for outlier detection can not be used for soil moisture data sets. The difficulty in detecting erroneous data is to differentiate between a severe rise of soil moisture due to precipitation events and an arbitrary sensor malfunction that results in an unpredictable value variation. Therefore some correct measurements are marked as erroneous observations, when automated quality control procedures are applied. The goal is to find a balance between "under flagged" erroneous data sets and "over flagged" correct data sets. To exclude a correct data set based on the quality assessment might seem unadvised, but when validating soil moisture data retrieved from satellite data one erroneous measurement does more harm than one missing correct observation.

2.2 Soil Moisture Sensors

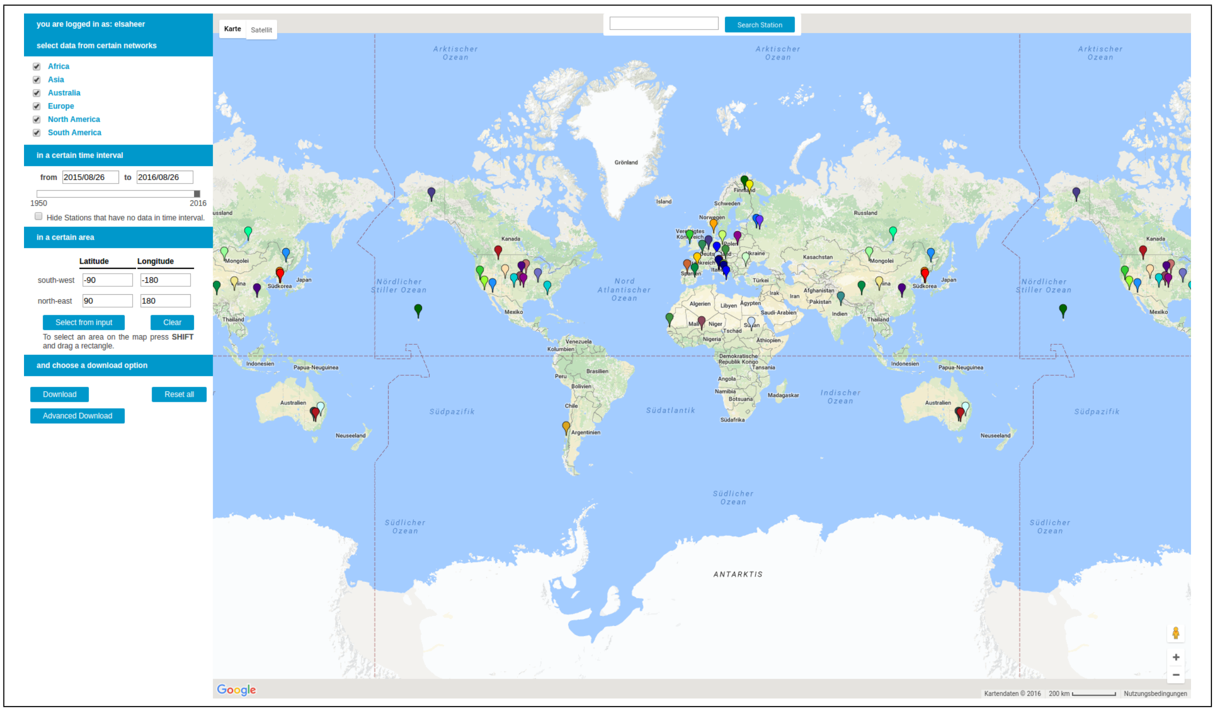
Soil moisture data is retrieved by many different methods. Hence, data sets measured with different sensors, show various acts of behavior. Additionally, the accuracy of in situ soil moisture sensors highly varies, depending on the method and quality of a sensor. There are different techniques of soil moisture retrieval introduced in [11], e.g. gravimetric, electromagnetic and remote sensing. Very accurate results can be achieved by gravimetric measurements. A soil sample is removed and weigh before and after a drying process. While the calculation is very simple and accurate, this method cannot be used for automated hourly measurements. That is why electromagnetic techniques are mostly common. An electromagnetic sensor is installed and provides periodical measurements. Electromagnetic sensors are distinguished into resistive and capacitive sensors and Time-Domain Reflectometer (TDR). With a resistive sensor the resistivity is measured, which depends on the moisture concentration as well as the ion concentration. Therefore a accurate calibration is essential, which is not constant. For the capacitive sensor two electrodes are installed and the capacitive between those is measured, and the dielectric constant calculated. The dielectric constant is directly proportional, but not linear to the soil water content. A calibration, considering temperature as well as soil type is needed. A TDR determines the propagation of electromagnetic waves. The propagation parameters, such as velocity and attenuation, depend on soil moisture and electrical conductivity. A TDR is a very reliable source, independent of soil texture and temperature, but comes with high cost. The network COsmic ray Soil Moisture Observing System (COSMOS) included in the ISMN provides soil moisture data derived from cosmic ray, reflected by the soil and measured above ground [12] [13]. This type of measure results in very noisy data. Another ISMN network Plate Boundary Observation (PBO_H2O) uses the picked up signal of Global Positioning System (GPS) receivers to calculate soil moister data [5].

With remote sensing data, which covers continuously large areas, soil moisture can be derived. It is calibrated and validated with in situ data sets. Therefore a quality assessment to exclude suspicious data sets is important to assure appropriate data sets to correlate with.

In [4], where the spatial representation of in situ observations were discussed, also the relation between sensors and specific types of errors was analyzed. The results were not distinct enough to undermine the assumption.

2.3 ISMN - International Soil Moisture Network

The operation of the ISMN started in 2010. The purpose of the ISMN is to provide in situ soil moisture data from data collectors all over the world in one single data portal (Figure 2.2). The data portal includes a data viewer for each station to show plots of data set timeseries, and a download center to retrieve data sets in one harmonized format, Coordinated Universal Time (UTC) and consistent units [2].



you are logged in as: elsaheer

select data from certain networks

- ☑ Africa
- ☑ Asia
- ☑ Australia
- ☑ Europe
- ☑ North America
- ☑ South America

in a certain time interval

from 2015/08/26 to 2016/08/26

1950 2016

□ Hide Stations that have no data in time interval.

in a certain area

	Latitude	Longitude
south-west	-90	-180
north-east	90	180

To select an area on the map press **SHIFT** and drag a rectangle.

and choose a download option

Figure 2.2: ISMN data portal

Nearly two thousand users are registered with the ISMN. All data sets of the ISMN are available to registered users, who also have the opportunity to share their experience with the ISMN data sets with other users and data providers in an online forum.

Several variables can be integrated in the ISMN, if the network provider measured and shared these data sets. Additional to soil moisture data the following variables can be included in the ISMN database:

- air temperature (ta)
- precipitation (p)

- snow depth (sd)
- snow water equivalent (sweq)
- soil suction (su)
- soil temperature (ts)
- surface temperature (tsf)

All data sets of a single data provider measured at various in situ stations are combined to a so called network. Each network includes a specific amount of stations, between one and over four hundred, and includes data sets in various time ranges of different variables in varying depth layers (Table 2.1).

Table 2.1: ISMN networks

network	country	stations	variables (layer)	time range
AACES	Australia	49	p(1) sm(3) ts(5)	2005/05/09 -2010/09/24
AMMA-CATCH	Benin, Niger, Mali	7	sm(12)	2006/01/01 -2014/12/31
ARM	USA	29	p(1) sm(10) ta(1) ts(10)	1993/06/29 -2015/03/26
AWDN	USA	50	sm(4)	1997/12/31 -2010/12/30
BIEBRZA_S-1	Poland	30	p(1) sm(4) ta(1) ts(4)	2015/04/23 -2016/06/24
BNZ-LTER	Alaska	12	p(1) sd(1) sm(4) sweq(1) ta(6) ts(9) tsf(1)	1988/06/01 -2013/01/01
CALABRIA	Italy	5	p(1) sm(3) ta(1)	2001/01/01 -2012/12/31
CAMPANIA	Italy	2	p(1) sm(1) ta(1)	2000/07/27 -2012/12/31
CARBOAFRICA	Sudan	1	p(1) sm(7) ta(1) ts(2)	2002/02/08 -2010/01/20
CHINA	China	40	sm(11)	1981/01/08 -1999/12/28
COSMOS	USA	109	sm(41)	2008/04/28 -2017/01/17
CTP_SMTMN	China	57	sm(4) ts(4)	2010/08/01 -2013/01/01
DAHRA	Senegal	1	p(1) sm(5) ta(1) ts(6)	2002/07/04 -2016/01/01
FLUXNET-AMERIFLUX	USA	2	p(1) sm(8) ta(1) ts(5)	2000/10/22 -2013/01/01
FMI	Finland	27	sm(7) ta(2) ts(7)	2007/01/25 -2017/01/14

GTK	Finland	7	sm(5) ta(1) ts(6)	2001/05/16 -2012/05/29
HOBE	Denmark	32	sm(3) ts(3)	2009/09/08 -2014/02/03
HSC_SELMACHEON	Korea	1	sm(1)	2008/01/01 -2009/01/01
HYDROL-NET_PERUGIA	Italy	2	p(1) sm(4) ta(1) ts(2)	2010/01/01 -2013/12/31
HYU_CHEONGMICHEON	Korea	1	sm(1)	2011/08/25 -2011/09/20
ICN	USA	19	p(1) sm(11) ts(6)	1983/01/03 -2010/11/21
IIT_KANPUR	India	1	sm(4)	2011/06/16 -2012/11/22
IOWA	USA	6	sm(12)	1972/04/04 -1994/11/15
iRON	USA	6	p(1) sm(3) ta(1) ts(1)	2012/08/21 -2016/01/01
KHOREZM	Uzbekistan	7	sm(1) ta(1) ts(5) tsf(1)	2010/04/17 -2011/09/10
LAB-net	Chile	1	p(1) sm(2) ta(1) ts(1)	2014/07/18 -2016/01/08
MAQU	China	20	sm(1) ts(1)	2008/06/30 -2010/07/31
METEROBS	Italy	1	sm(5)	2011/10/23 -2012/05/09
MOL-RAO	Germany	2	p(1) sm(9) ta(2) ts(12)	2003/01/01 -2014/01/01
MONGOLIA	Mongolia	44	sm(10)	1964/04/08 -2002/10/18
ORACLE	France	6	p(1) sm(26) ta(2) ts(2)	1985/10/18 -2013/09/09
OZNET	Australia	38	p(2) sm(7) su(5) ts(6)	2001/09/12 -2011/05/31
PBO_H2O	USA	161	p(1) sd(1) sm(1) ta(1)	2004/09/27 -2017/01/14
REMEDHUS	Spain	24	sm(1) ts(1)	2005/03/15 -2017/01/01
RSMN	Romania	20	p(1) sm(1) ta(1) ts(1)	2014/04/09 -2016/12/31
RUSWET-AGRO	Former Soviet Union	212	sm(2)	1958/04/08 -2002/06/28
RUSWET-GRASS	Former Soviet Union	122	sm(2)	1952/06/08 -1985/12/28
RUSWET-VALDAI	Former Soviet Union	3	p(1) sm(3) ta(1) ts(3)	1960/01/15 -1990/12/15
SASMAS	Australia	14	sm(2) ts(1)	2005/12/31 -2007/12/31
SCAN	USA	232	p(1) sd(1) sm(25) sweq(1) ta(1) ts(25)	1996/01/01 -2017/01/17
SKKU	Korea	5	sm(4)	2014/05/08 -2014/11/19
SMOSMANIA	France	21	sm(4) ts(4)	2007/01/01 -2016/01/01
SNOTEL	USA	441	p(1) sd(1) sm(16) sweq(1) ta(1) ts(16)	1980/10/01 -2017/01/17
SOILSCAPE	USA	171	sm(28) ts(1)	2011/08/03 -2016/11/01

SWEX_POLAND	Poland	6	p(1) sm(10) ts(8)	2000/01/01 -2013/05/06
SW-WHU	China	7	p(1) sm(1) ta(1) ts(1)	2014/01/12 -2015/06/03
TERENO	Germany	5	p(1) sm(3) ta(1) ts(3)	2009/12/31 -2017/01/05
UDC_SMOS	Germany	11	sm(5)	2007/11/08 -2011/11/18
UMBRIA	Italy	13	p(1) sm(5) ta(1)	2002/10/09 -2014/08/01
UMSUOL	Italy	1	sm(7)	2009/06/12 -2010/09/30
USCRN	USA	115	p(1) sm(5) ta(1) ts(5) tsf(1)	2000/11/15 -2017/01/16
USDA-ARS	USA	4	sm(1) ts(1)	2002/06/01 -2009/07/31
VAS	Spain	3	sm(1) ta(2) ts(10)	2010/01/01 -2012/01/01
WEGENERNET	Austria	12	p(1) sm(2) ta(1) ts(2)	2007/01/01 -2016/12/22
WSMN	UK	8	sm(3) ts(3)	2011/09/02 -2016/02/29
Total: 55		2226		1952/06/08 -2017/01/17

Since the start of the ISMN 55 data providers shared their data. In total there are 2226 in situ stations, where soil moisture and optionally other variables in different depth layers are measured. These data sets come in varying file formats. They not only differ in format, time and units, which is edited to be consistent by the ISMN processing chain, but also use different sensor types, which leads to diverging data behaviors. Data sets in the ISMN are normally integrated as hourly data. If the measurements are more frequent than once per hour only the observation, nearest to the full hour is taken. In rare cases the measurement frequency is lower than one hour. (e.g. daily data sets from the network PBO_H2O)

3 Existing Methods

The automated quality control system of the ISMN is very innovative and advanced. It is not common use to provide data sets with an appropriate quality assessment. Only 5 of the 55 data providers of the ISMN have their own quality control system. Also the North American Soil Moisture Databse (NASMD), a project similar to the ISMN for the United States only, has a flagging systems for which they refer to the methods of the ISMN [9].

The automated quality control procedure of the ISMN is a system, within the data processing chain, that checks every data value that is integrated in the ISMN database for its correctness, or rather its reasonableness, and appends one or more appropriate letter coded quality flag to the measurement. The flags are letter coded according to the Code of Practice (CEOP) standard
(https://www.eol.ucar.edu/projects/ceop/dm/documents/refdata_report/).

The ISMN quality flags are parted in three categories: exceeding boundary values, geophysical consistency checks and a spectrum based approach (Table 3.1).

Table 3.1: ISMN quality flag categories

flag	category
C01 - C03	reported value exceeds output format field size
D01 - D05	questionable/dubious - geophysical based
D06 - D10	questionable/dubious - spectrum based

The C flags trigger, if a value exceeds a boundary minimum or maximum. These boundaries are defined for each variable respectively and are rather tolerant, since the ISMN includes data from all around the world, measured at different land cover and climate classes (Table 3.2).

The D flags are only implemented for soil moisture values. Geophysical reasons are defined as inconsistencies with other variables, like soil temperature or precipitation.

Table 3.2: ISMN quality flags that represent an exceeding boundary error

variable	flag	condition
air temperature	C01	air temperature < -60°C
	C02	air temperature > 60°C
precipitation	C01	precipitation < 0 mm/h
	C02	precipitation > 100 mm/h
snow depth	C01	snow depth < 0 mm
snow water equivalent	C01	snow water equivalent < 0mm
soil moisture	C01	soil moisture < 0.0 m^3/m^3
	C02	soil moisture > 0.6 m^3/m^3
	C03	soil moisture > staturation point (derived from HWSD parameter values)
soil suction	C01	soil suction < 0 kPa
	C02	soil suction > 2500 kPa
soil temperature	C01	soil temperature < -60°C
	C02	soil temperature > 60°C
surface temperature	C01	surface temperature < -60°C
	C02	surface temperature > 60°C

Those data sets are retrieved from either the same in situ station or the Global Land Data Assimilation System (GLDAS) data sets, a data model provided by National Space Agency (NASA). The focus of this thesis will be on algorithms for the detection of spectrum based data errors, which were specifically developed for the ISMN. Spectrum based flags are applied, if the soil moisture timeseries itself shows dubious effects at one observation or over a series of measurements (Table 3.3). A certain time range of the soil moisture timeseries and its first and second derivative are therefore examined for a certain suspicious behavior. These methods will be analyzed for their reliability, meaning if they miss erroneous observations or falsely flag correct measurements.

Only five data providers, contributing to the ISMN, provide their own automated flagging system and deliver their data sets with a corresponding flag for each data value. These quality control systems are called original quality flags for the ISMN. They also use letter or number coded quality flags for their assessment. These networks will be introduced and their provided quality flags will be compared to those of the ISMN. The goal is to learn from those other system and integrate their ideas into the

Table 3.3: ISMN quality flags that represent a geophysical or spectrum based error

variable	flag	condition
soil moisture	D01	in situ soil temperature < 0°C at corresponding depth layer
	D02	in situ air temperature < 0°C
	D03	GLDAS soil temperature < 0°C at corresponding depth layer
	D04	soil moisture shows peaks without in situ precipitation event in preceding 24 hours
	D05	soil moisture shows peaks without GLDAS precipitation event in preceding 24 hours
	D06	a spike in soil moisture spectrum
	D07	a negative jump in soil moisture spectrum
	D08	a positive jump in soil moisture spectrum
	D09	low constant values (minimum 12 hours) in soil moisture spectrum
	D10	saturated plateau (minimum 12 hours) in soil moisture spectrum

algorithms of the ISMN, if they are a valuable addition.

3.1 Flagging Statistics for the ISMN

Data sets that are shared with the ISMN by different network providers go through a long processing chain to harmonize the format. At the end of this chain each value receives a letter coded quality flag. For each flag several conditions have to be met to add the specific letter to the value. If no flag is triggered the value receives the flag G for 'Good'. The behavior of the in situ soil moisture data sets and their quality vary highly per network, due to the use of different sensors and various climate and land cover classes. Therefore the percentages of triggered quality flags per network are also very diverse (Figure 3.1).

The table shows that exceeding boundary and geophysical flags occur more often than spectrum based flags. Since those errors rely on facts, e.g. air temperature is below zero degrees or they exceed a boundary value, they are much easier detected than errors in the spectrum. Spectrum based errors on the other hand are detected with conditions that have to fit every occurrence of the error in various spectra. An important goal for the improvement of the existing algorithms for the spectrum based flags is, to improve the detection rate of the algorithms. For the validation of satellite data, for example, a bad observation that was not excluded is much more severe for a correlation than a good measurement that is missing. Therefore also more over flagged correct data sets can be accepted to improve the detection rate of erroneous observations.

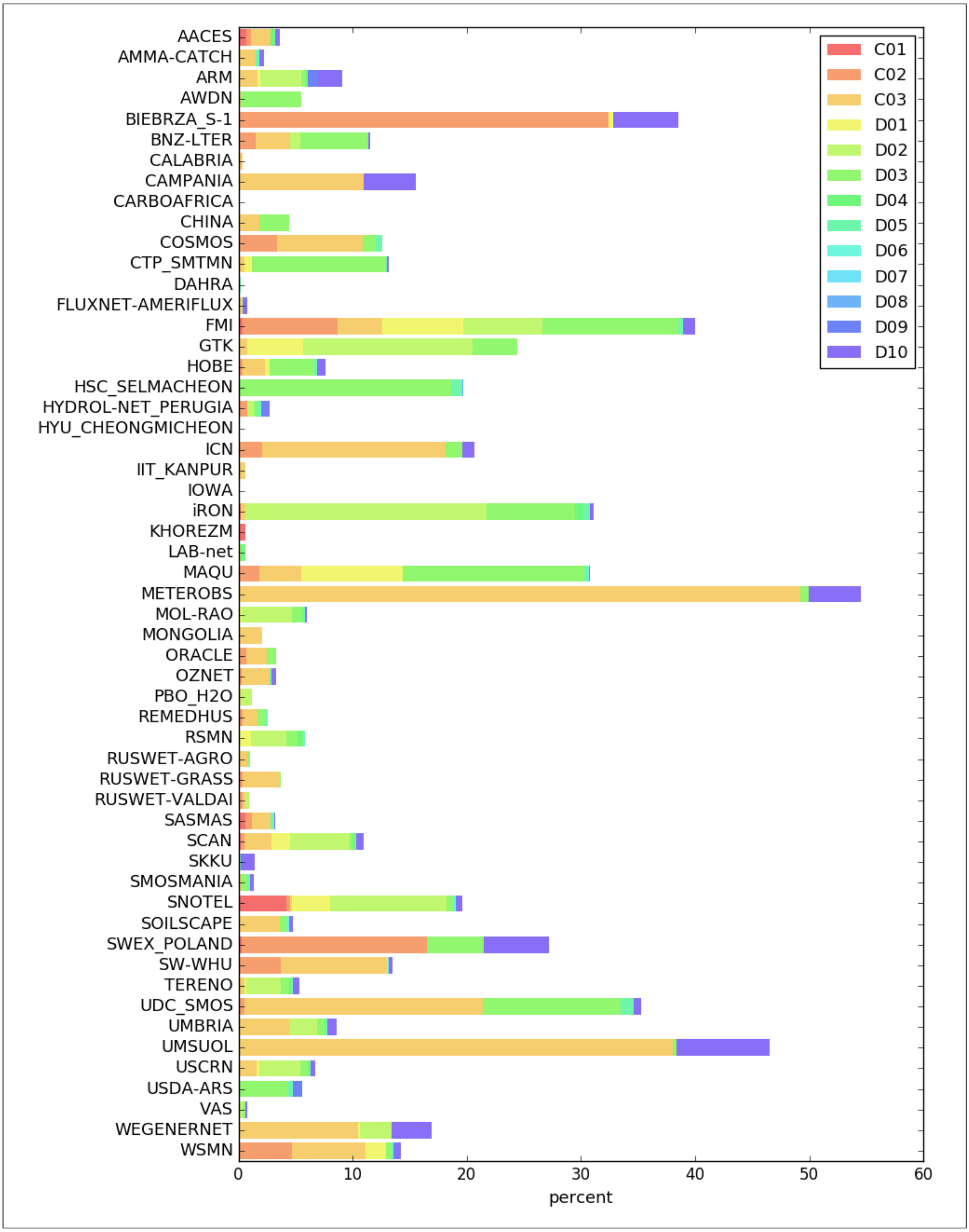


Figure 3.1: ISMN flagging statistic for each network

3.2 ISMN Algorithms for the Detection of Spectrum Based Errors

Spectrum based errors are sudden unnatural changes in a soil moisture timeseries. Those errors are detected by examining the shape of a soil moisture timeseries and its first and second derivative. The spectrum based flags are developed for hourly data sets, since the integrated data sets in the ISMN are mostly hourly.

3.2.1 Savitzky-Golay Filter

The savitzky-golay filter was used to calculate the first and second derivative of a soil moisture timeseries [7]. The savitzky-golay filter calculates a regression of discrete data values to provide a smoothed data set. The savitzky-golay filter does not cut off high frequencies, but considers all data values in its calculation. It takes data observations, the range of data points that should be considered for the filtering and the degree of the polynomial as input values. For the purpose of the ISMN also the optional argument of a derivative was used. For spikes and breaks the filter range is set to three observations and for plateaus it is set to 25. In all cases a polynomial of degree two was chosen.

3.2.2 Spikes

A spike is an event that lasts only for one measurement and differs noticeable from the values before and after. The outlier can point in a positive or negative direction and mostly represents a sensor malfunction or energy supply shortage. The first derivative shows a positive and a negative peak and the second derivative a very high peak and two smaller peaks before and after. For a positive spike the big peak is negative and the smaller ones are positive. In the ideal case the big peak is twice as high as the smaller ones (Figure 3.2). In the figures x_t represents the soil moisture timeseries, the gray x'_t the actual first derivative, $\tilde{x}'_t$ the rounded first derivative that is used for the calculations and x''_t the second derivative.

To mark an observation as a spike three conditions have to be met.

1. The increase or decrease of soil moisture amounts to 15%, which typically rep-

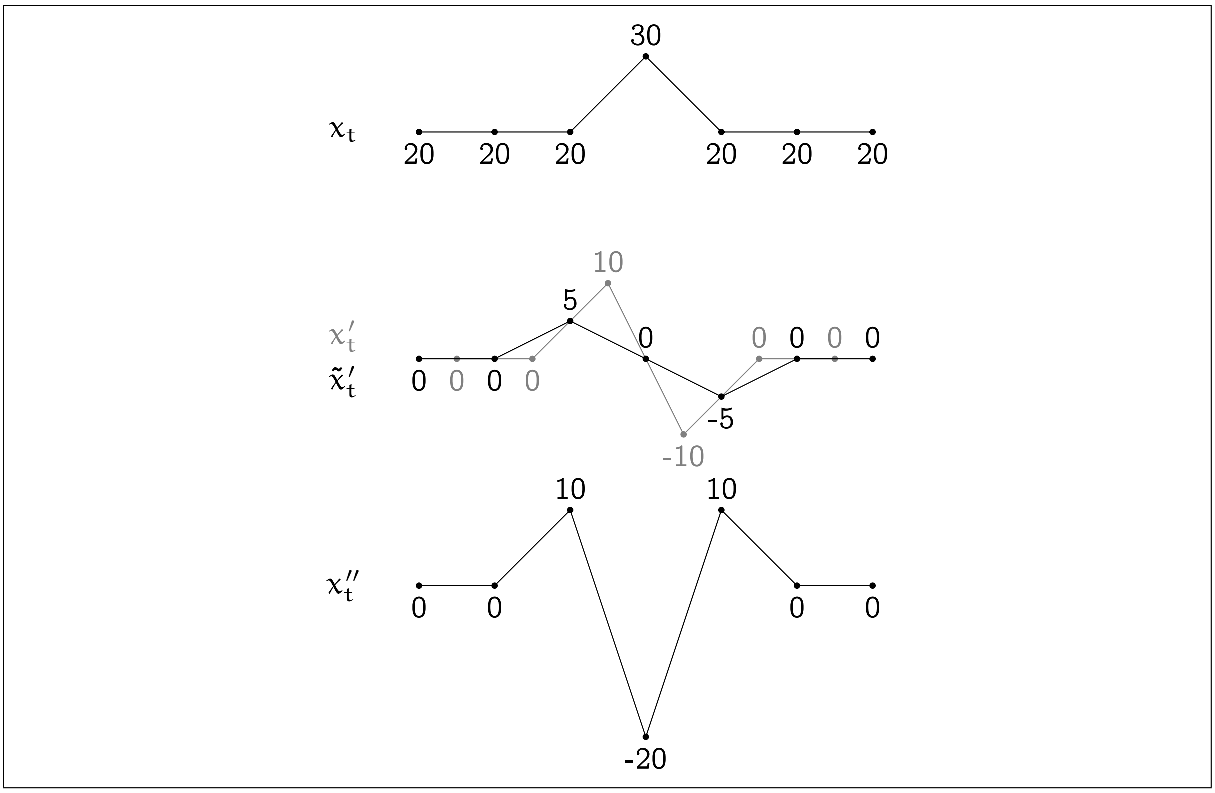


Figure 3.2: data set x_t with a spike and its first derivative $\tilde{x}'_t$ and second derivative x''_t

resents three times the maximal sensor uncertainty.

$$\frac{x_t}{x_{t-1}} > 1.15 \quad \text{or} \quad \frac{x_t}{x_{t-1}} < 0.85 \tag{3.1}$$

2. To distinguish between a soil moisture rise due to a precipitation event and a spike, the values before and after the spike have to be approximately the same. A positive/negative spike results in a very high negative/positive peak and a small positive/negative peak before and after, in the second derivative. The small peaks are compared to be nearly equal in the second condition.

$$0.8 < \left| \frac{x''_{t-1}}{x''_{t+1}} \right| < 1.2 \quad \text{with} \quad x''_{t+1} \neq 0 \tag{3.2}$$

3. To avoid over flagging of noisy data the third condition requests the coefficient of variation to be less than one. This coefficient is a typical proportion, used as

an order for a relative spreading of data values.

$$\left|\frac{\sigma^2([x_{t-12}, x_{t+12}]\backslash x_t)}{\mu([x_{t-12}, x_{t+12}]\backslash x_t)}\right| \quad < \quad 1 \tag{3.3}$$

The observation x_t is flagged as a spike if all three conditions are met.

3.2.3 Breaks

A break is a noticeable rise or fall in a soil moisture timeseries, compared to the measurement before. The spectrum before and after the break would seem continuous without the erroneous event.

Negative Breaks

The negative break shows a peak at the first derivative and alternating peaks in the second derivative (Figure 3.3).

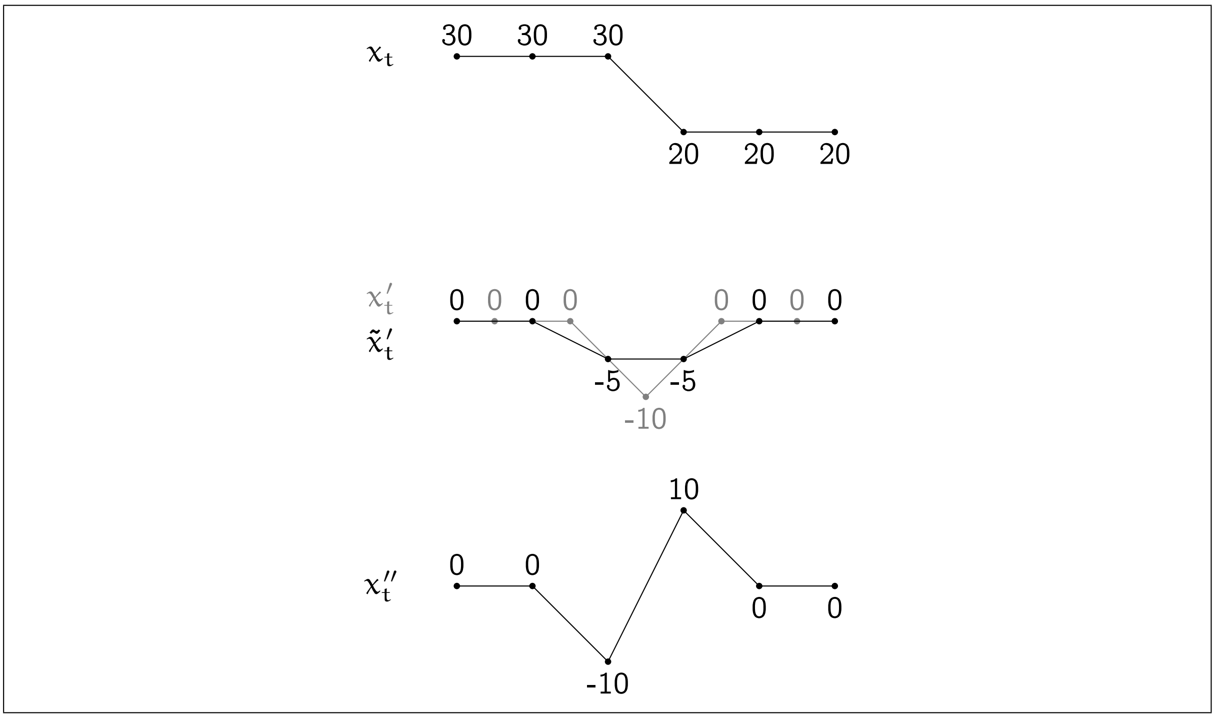


Figure 3.3: data set x_t with a negative break and its first derivative $\tilde{x}'_t$ and second derivative x''_t

1. The first condition states that the soil moisture drop has to be at least 10 percent. The absolute difference in m^3/m^3 between the values was defined to avoid over flagging at low soil moisture readings.

$$\left|\frac{x_t - x_{t-1}}{x_t}\right| > 0.1 \quad \text{with} \quad x_t \neq 0 \quad \text{and} \quad |x_t - x_{t-1}| > 0.01 \tag{3.4}$$

2. The second condition represents the negative peak in the first derivative, which was defined to be at least ten times smaller than the 12 hour spectrum before and after the break.

$$x'_t < -10 \cdot \frac{1}{25} \cdot \left|\sum_{k=-12}^{12} x'_{t+k}\right| \tag{3.5}$$

3. The third condition includes characteristics of the second derivative. The values at the break and before the break should be equal and very high compared to the value after the break.

$$\left|\frac{x''_t}{x''_{t-1}}\right| = 1 \quad \text{with} \quad x''_{t-1} \neq 0 \quad \text{and} \quad \left|\frac{x''_t}{x''_{t+1}}\right| > 10 \quad \text{with} \quad x''_{t+1} \neq 0 \tag{3.6}$$

Positive Breaks

The positive break is the complement to the negative break (Figure 3.4), but it is much harder to detect due to the natural characteristics of soil moisture readings. Even more difficult is the prevention of over flagging precipitation events. Nevertheless, the formulas are similar to those of the negative break.

1. Analog to the negative break the soil moisture rise must be higher than 10 percent.

$$\left|\frac{x_t - x_{t-1}}{x_t}\right| > 0.1 \quad \text{with} \quad x_t \neq 0 \quad \text{and} \quad |x_t - x_{t-1}| > 0.01 \tag{3.7}$$

2. For the positive break the first derivative at the break has to be at least ten times bigger than the spectrum before and after.

$$x'_t > 10 \cdot \frac{1}{25} \cdot \left|\sum_{k=-12}^{12} x'_{t+k}\right| \tag{3.8}$$

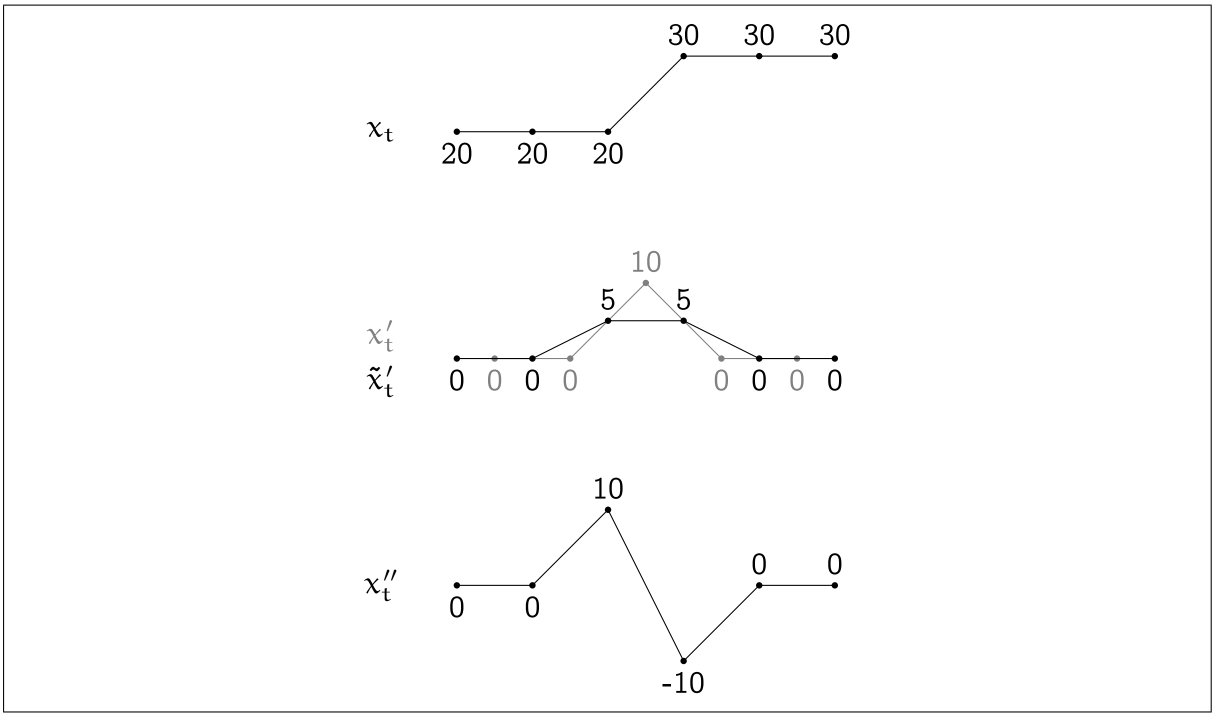


Figure 3.4: data set x_t with a positive break and its first derivative $\tilde{x}'_t$ and second derivative x''_t

3. The last equation is the same as for the negative break.

$$\left|\frac{x''_t}{x''_{t-1}}\right| = 1 \quad \text{with} \quad x''_{t-1} \neq 0 \quad \text{and} \quad \left|\frac{x''_t}{x''_{t+1}}\right| > 10 \quad \text{with} \quad x''_{t+1} \neq 0 \tag{3.9}$$

In most of the cases one cannot tell whether the values before or after the break are suspicious. Therefore only the break itself is flagged.

3.2.4 Plateaus

Plateaus are defined as ongoing relatively constant values. There are low level and high level plateaus, which vary in their characteristics and therefore the algorithms are different. For the error detection of the ISMN a plateau was defined to last at least 12 hours, even though the figures suggest otherwise for simplification.

Low Constant Values

Low constant values are primarily the result of an energy supply shortage. They follow a negative break with unfitting low soil moisture readings (Figure 3.5). There are only two conditions for a low level plateau.

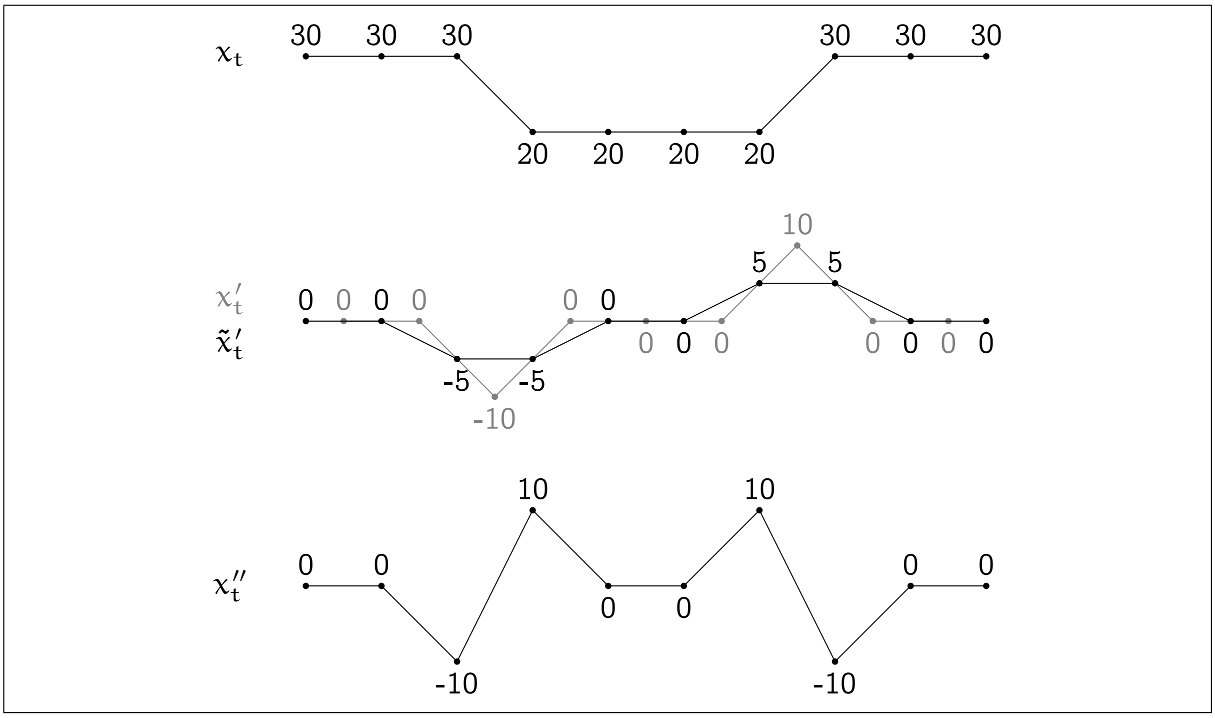


Figure 3.5: data set x_t with a low level plateau and its first derivative $\tilde{x}'_t$ and second derivative x''_t

1. A negative break was detected as the start of the plateau.
2. The coefficient of variation is again used to guarantee a very constant progression of the plateau. It must be smaller than 1.

$$\left|\frac{\sigma^2([x_t, x_{t+n}])}{\mu([x_t, x_{t+n}])}\right| < 1 \quad \text{with} \quad n \geqslant 12 \quad \text{and} \quad t = t_{pl\_start} \tag{3.10}$$

Saturated Plateaus

Saturated plateaus are often the result of a sensor malfunction due to values above the saturation point (Figure 3.6). Again, a plateau was defined to last at least 12 hours to avoid over flagging.

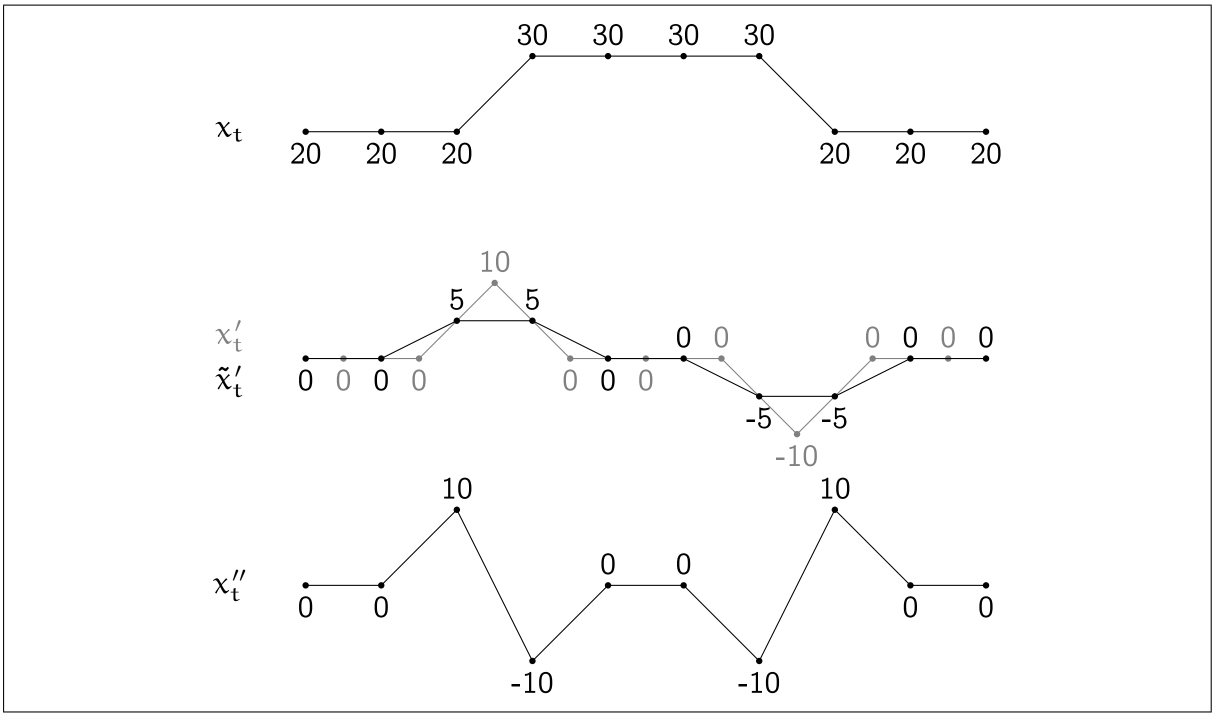


Figure 3.6: data set x_t with a saturated plateau and its first derivative $\tilde{x}'_t$ and second derivative x''_t

1. The values of the saturated plateau must only vary very little. As boundary 1 percent of the the typical sensor variation at 5 percent, was chosen, which results to 0.05 percent volume soil moisture variation.

$$\sigma^2([x_{t-n}, x_{t+n}]) < 0.05 \quad \text{with} \quad n \geqslant 6 \quad \text{and} \quad t = t_{pl\_start} \tag{3.11}$$

2. A saturated plateau often appears after a heavy precipitation event. Therefore it follows a soil moisture rise and is followed by a drop of soil moisture, when the sensor reacts to lower values again. These rises and drops do not have to be exactly at the beginning or end of the plateau, because the sensor reaction may stop at a time stamp after a big rise and start again slowly afterwards. They are defined to happen in a spectrum 12 hours before or after the beginning and the end of the plateau.

$$\max([x'_{t-n-12}, x'_{t-n+12}]) \geqslant 0.25 \quad \text{with} \quad t = t_{pl\_start} \tag{3.12}$$

$$\min([x'_{t-n-12}, x'_{t-n+12}]) \leqslant 0 \quad \text{with} \quad t = t_{pl\_end} \tag{3.13}$$

3. Since saturated plateaus only appear at very high soil moisture values the third condition states that the mean value of the plateau must be in the upper 5 percent of all observations of the timeseries.

$$\mu([x_{pl\_start}, x_{pl\_end}]) > \max([x_{t_0}, x_{t_{end}}] \setminus \{x_t > 100\}) \cdot 0.95 \tag{3.14}$$

3.3 Problems with the Existing ISMN Quality Flags

Since the start of the ISMN, its database grew very quickly. More and more data providers want to make their data available for other scientific research through the ISMN. Therefore also more erroneous measurements were identified, since the development of the automated quality control system in 2013. For every spectrum based flag an under flagging of erroneous measurements as well as an over flagging of correct observations was identified. In many cases various thresholds of conditions are too tight and avoid the flagging of erroneous data values (Figure 3.7).

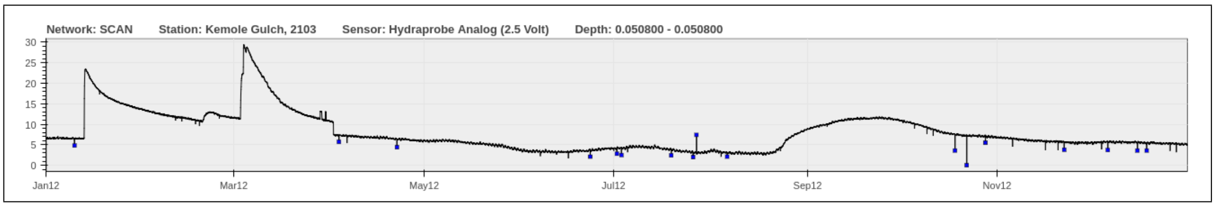


Figure 3.7: under flagged data set due to tight threshold conditions

3.3.1 Spikes

The threshold of the first condition (Equation 3.1), which compares the relation between the spike and the values before and after respectively, and the second condition (Equation 3.2) for the second deviation are very tight. Thus, many spikes are not flagged as such, because the specific values cannot meet those boundaries. A derestriction to looser values of those boundaries would result in an over flagging of correct values. Thus, new additional conditions would have to be introduced.

3.3.2 Breaks

A reason why some breaks are still not detected is the third condition (Equation 3.6), where x''_{t+1} is the denominator, which is zero in the ideal case. An inversion of this division would fix the problem. In equation 3.5 the rounding of the first derivative was not considered. x'_t and x'_{t-1} always have to be treated equally. Otherwise the negative breaks are well detected by the existing algorithms. A very complex problem is the over flagging of correct data sets with positive breaks. While a sudden drop in a soil moisture spectrum is a very suspicious event, a positive rise occurs naturally due to a precipitation event or snow melting process. More conditions especially for positive breaks are necessary. Additionally, the thresholds of the first conditions (Equation 3.4 and 3.7) can be loosened to avoid the under flagging of erroneous observations.

3.3.3 Plateaus

Plateaus are under flagged by the existing methods. A negative break as the start of constant values is the first condition of a low level plateau. Hence, if the detection rate of negative breaks would be improved, also low level plateaus that failed only this first condition would be flagged. A rise of soil moisture values due to natural events and a following slow drying process is often mistaken for a positive plateau. Therefore the conditions are chosen very tight and many saturated plateaus are not detected. A better balance between under and over flagging has to be found.

3.3.4 Errors that are not Covered by the ISMN Quality Flags

Other errors than those described by the ISMN quality flags were visually identified since their development. Data values before or after a period of missing values are often outliers. Since there is no spectrum available during the interrupted time, the erroneous value cannot be detected with the existing methods. Constant values in a soil moisture spectrum are very suspicious, since it is effected by many factors. Constant values are mostly the result of frozen soil or a battery shortage. The former should be covered by the geophysical flags that respond if the air or soil temperature is below zero degree. Those often do not work in spring when the sensor needs time to recover from the freezing soil. A very specific erroneous behavior of soil moisture values that is not covered by the ISMN quality control methods, is jumping between two data values over a specific time period. Finally, if a data value lies within a spectrum where many other data values before and after are flagged, it is very unlikely that the certain value is correct, since the sensor is malfunctioning. Often scattered values within an erroneous spectrum cannot be detected by algorithms, because of the arbitrary variation in such time periods.

3.4 Other Flagging Systems

3.4.1 Original Quality Flags of ISMN Networks

Flagging systems of ISMN network providers that are integrated in the ISMN data base are defined as original quality flags within the ISMN. Five of the ISMN networks provide their own quality assessment. The flags of all data sets of those networks were compared to those of the ISMN.

The ISMN quality control system has a higher detection rate than any of the providers. In the following this will be analyzed in detail. The reason for the differences is primarily that not every provider has a quality control procedure for every variable, but only for some of them. The available research material of the providers will be analyzed to find other explanations of the differences. Of special interest are those data values that are flagged by the providers and not by the ISMN to expand the quality control system of the ISMN.

Bonanza Creek Long-Term Ecological Research (BNZ-LTER)

The BNZ-LTER network is located in Alaska and consists of 12 station.

The operation started in 1987 and is managed by the Institute of Arctic Biology, University of Alaska Fairbanks. The data sets included in the ISMN range from 1988 to 2012. The BNZ-LTER project provides experimental and observational research designed to understand the dynamics, resilience, and vulnerability of Alaska's boreal forest ecosystems. Their also letter coded flagging system includes four flags:

- G - Good,
- Q - Questionable,
- M - Missing, and

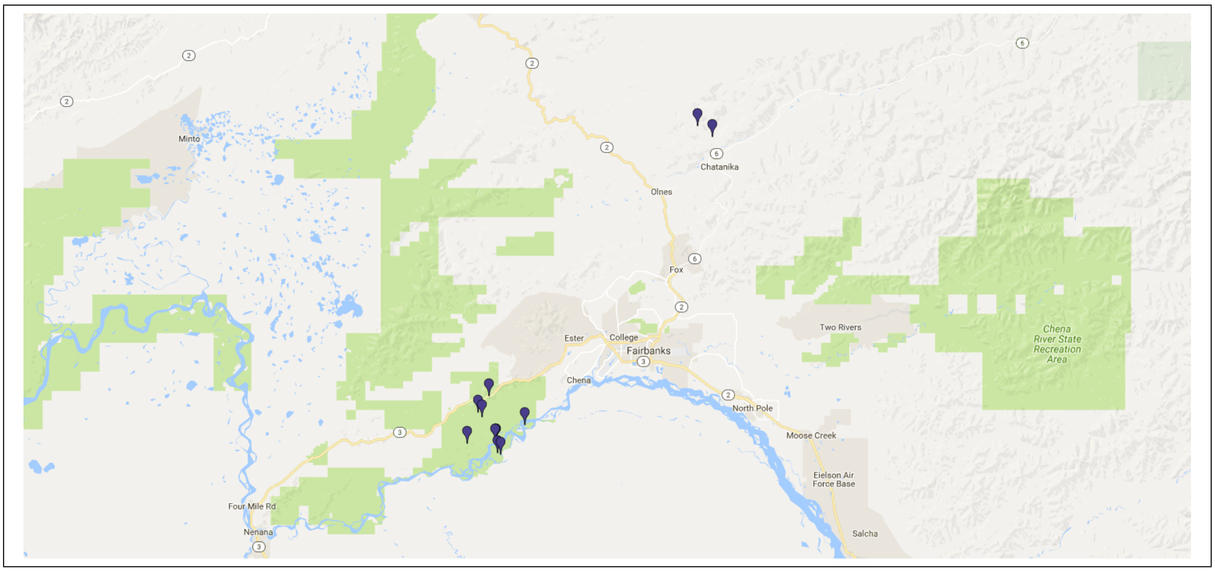


Figure 3.8: stations of the BNZ-LTER network

- E - EstimatedMissing.

Unfortunately no publication is available for a detailed description of their flagging system. It could not be determined, if the flags are designated for all variables. Table 3.4 shows that not a single data value, except for surface temperature, was flagged.

Table 3.4: comparison of ISMN and BNZ-LTER flagging statistics

		flagged by ... [absolute number | % of datasets]											
variable	datasets	either		both		ismn		provider		ismn only		provider only	
p	389,454	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
sd	185,965	55	0.03	0	0.00	55	0.03	0	0.00	55	0.03	0	0.00
sm	4,058,604	1,752,304	43.18	0	0.00	1,752,304	43.18	0	0.00	1,752,304	43.18	0	0.00
sweq	119,700	194	0.16	0	0.00	194	0.16	0	0.00	194	0.16	0	0.00
ta	3,553,801	12	0.00	0	0.00	12	0.00	0	0.00	12	0.00	0	0.00
ts	10,896,930	8	0.00	0	0.00	1	0.00	7	0.00	1	0.00	7	0.00
tsf	345,972	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
total	19,550,426	1,752,573	8.96	0	0.00	1,752,566	8.96	7	0.00	1,752,566	8.96	7	0.00

The network however shall be mentioned for completeness anyway and hopefully information about their system will be available in the future.

Lindenberg Meteorological Observatory - Richard Aßmann Observatory (MOL-RAO)

The MOL-RAO network is located in Germany and consists of two stations, one located in grass land and the other in a pine forest (Figure 3.9).

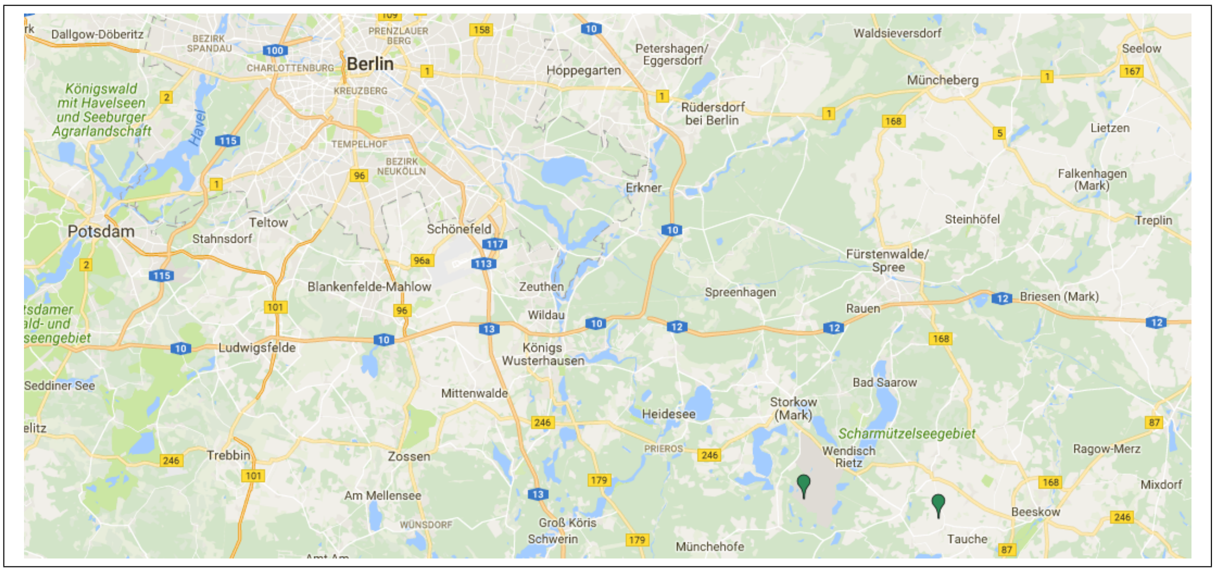


Figure 3.9: stations of the MOL-RAO network

The operation started in 2003 by the German Meteorological Service. The data sets included in the ISMN range from 2003 to 2013. For the network MOL-RAO there exist six different, also letter coded, quality flags:

- G - Good,
- B - Bad,
- D - Dubious,
- M - Missing,
- I - Interpolated, and
- U - Unchecked.

The original measurements, observed with a sampling frequency of 10 minutes, are averaged to a 30 minute interval for all variables. The flags of those observations are then combined to one flag depending on their amount of occurrences defined by a look up table.

Similar to the C flag methods of the ISMN, data values are checked for exceeding a certain threshold. If they do they are transformed to Not a Number (NaN) values (-9999.99) and get the missing flag M. The two stations are not far from each other. Hence, the boundaries are tighter than those of the ISMN, because they soil and climate properties are as diverse as for all ISMN stations. For soil moisture values the boundaries are from 3% to 50%.

Flag D of the MOL-RAO quality system is automatically appended if auxiliary measurements or physical arguments are inconsistent with the measured variable. There is a test for air temperature, soil temperature and soil moisture. For air temperature the measurements of two different sensors are compared and if the exceed a difference of 1 K the measurements fail the test. The soil temperature values are compared to other depths and flagged if the difference is higher than a certain threshold depending on the depth difference. The soil moisture measurements are compared to the neighboring values in depth and time and only pass the test if they do not exceed 5% or 50% of the measured value. D flags are controlled manually and are either transformed to a G or B flag or in case of doubt left as a D flag. If a flag is not checked, an U flag is appended to the D flag.

The delivered measurements of the MOL-RAO network are every 30 minutes. Since the ISMN only includes hourly data, the flags may not always be optimal for the whole hour.

Table 3.5 shows that a lot less soil moisture values are flagged than by the ISMN, but also a lot of values only by the provider.

Table 3.5: comparison of ISMN and MOL-RAO flagging statistics

		flagged by ... [absolute number | % of datasets]											
variable	datasets	either		both		ismn		provider		ismn only		provider only	
p	189,485	8,632	4.56	0	0.00	0	0.00	8,632	4.56	0	0.00	8,632	4.56
sm	916,944	169,624	18.50	22,718	2.48	153,140	16.70	39,202	4.28	130,422	14.22	16,484	1.80
ta	192,651	1,204	0.62	0	0.00	0	0.00	1,204	0.62	0	0.00	1,204	0.62
ts	1,787,951	16,473	0.92	0	0.00	0	0.00	16,473	0.92	0	0.00	16,473	0.92
total	3,087,031	195,933	6.35	22,718	0.74	153,140	4.96	65,511	2.12	130,422	4.22	42,793	1.39

Analyzing those data sets showed that those are highly over flagged. A rise in soil moisture can easily be over absolute 5 % or relative 50 % after a precipitation event. All those rises are flagged as erroneous data even if they follow a natural event (Figure 3.10). The original quality flags are indicated by a vertical red line.

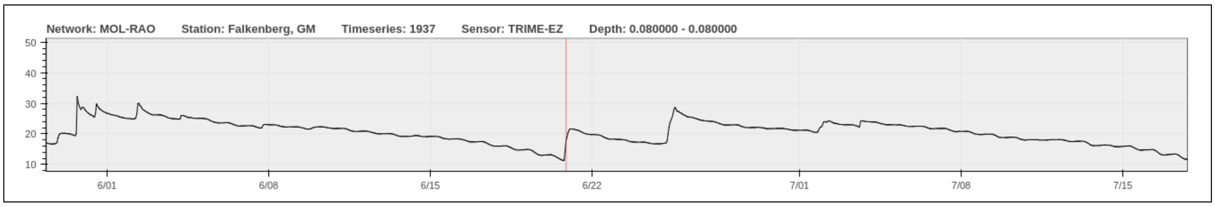


Figure 3.10: over flagged MOL-RAO data set

In the other direction however, it seems appropriate to flag such a high soil moisture drop.

SOIL moisture Sensing Controller And oPtimal Estimator (SOILSCAPE)

The SOILSCAPE network is located in the United States, including 136 stations, piled in the mid and western USA. The operation started in 2011 and is managed by the USC (University of Southern California) [6]. Their quality assessment includes six flags that are number and letter coded for soil moisture data observations:

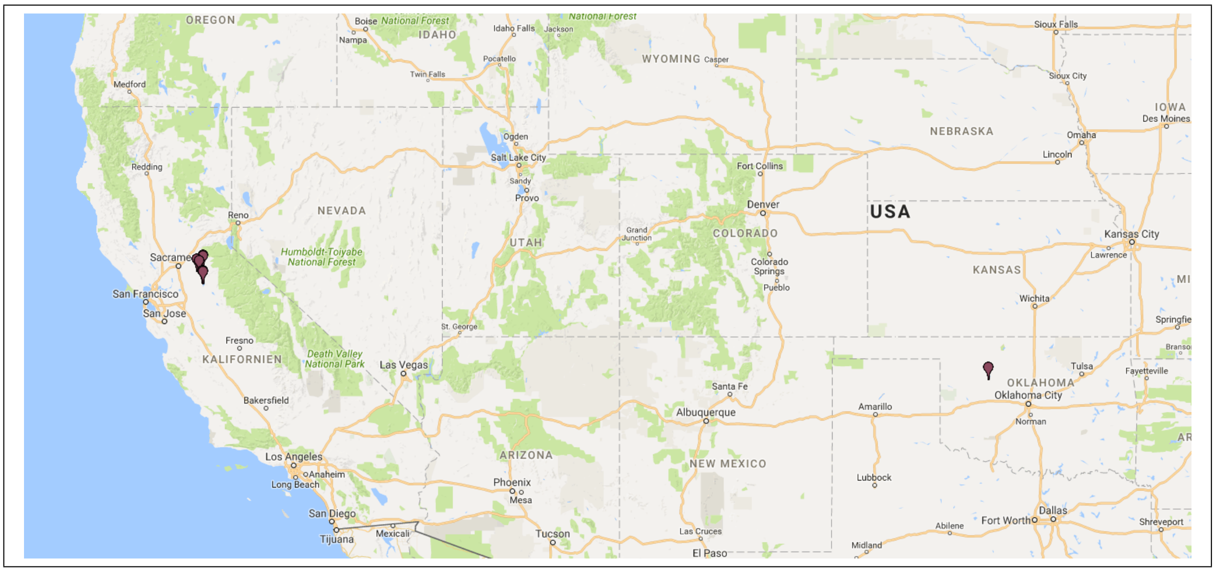


Figure 3.11: stations of the SOILSCAPE network

- 0 - (G) Good (Standard for all data),

- 1 - (D) Dubious (automatically flagged, spikes etc.),
- 2 - (I) Interpolated / Estimated,
- 3 - (B) Bad (Manually flagged),
- 4 - (M) Missing, and
- 5 - (C) Exceeds field size (Negative SM values, fixed at 0.1 percent).

Table 3.6: comparison of ISMN and SOILSCAPE flagging statistics

		flagged by ... [absolute number | % of datasets]											
variable	datasets	either		both		ismn		provider		ismn only		provider only	
sm	4,233,386	260,306	6.15	42	0.00	257,025	6.07	3,323	0.08	256,983	6.07	3,281	0.08
ts	84,279	194	0.23	0	0.00	194	0.23	0	0.00	194	0.23	0	0.00
total	4,317,665	260,500	6.03	42	0.00	257,219	5.96	3,323	0.08	257,177	5.96	3,281	0.08

For the network SOILSCAPE also no detailed information was available. An analyzing of the data sets showed that the soil moisture values were very under flagged, which is also indicated in Table 3.6. Where they have more success in flagging erroneous data than the ISMN is at continuous dubious data sets over a very long time period. Those series seem to be flagged manually, since they perfectly fit and the area and don't seem to be possibly detected by any algorithm.

TERENO

The Terrestrial Environmental Observatories (TERENO) network is located in Germany and consists of five stations [10]. Their flags are number coded:

- 4_2 - good or acceptable data,
- 4_15 - irregular data,
- 4_18 - isolated spike, and
- 4_19 - sensor is out of order or implausible data.

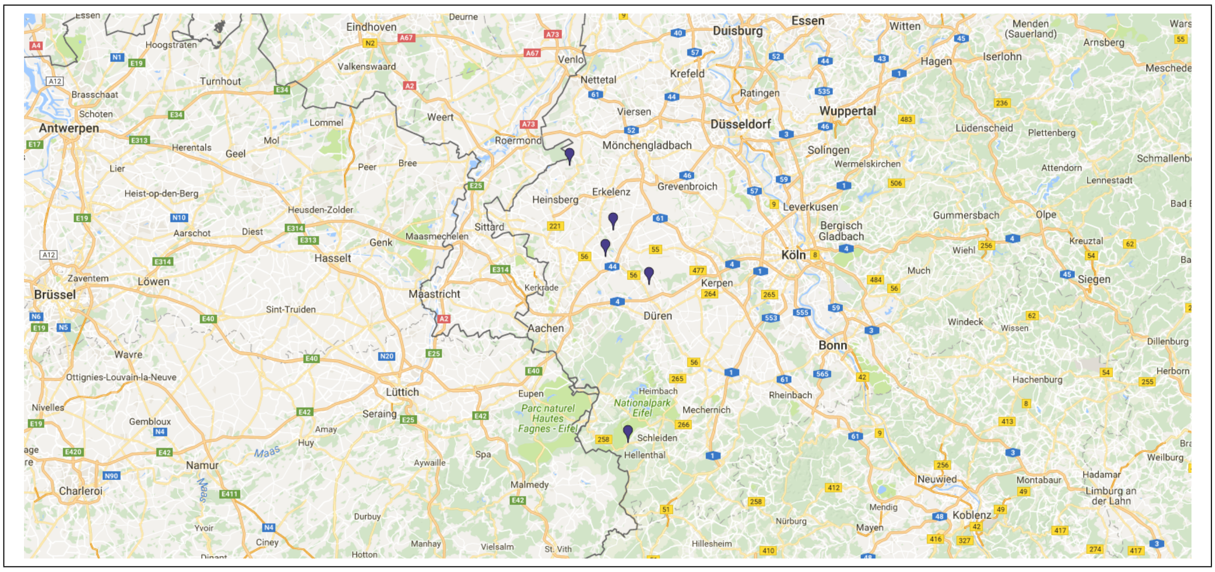


Figure 3.12: stations of the TERENO network

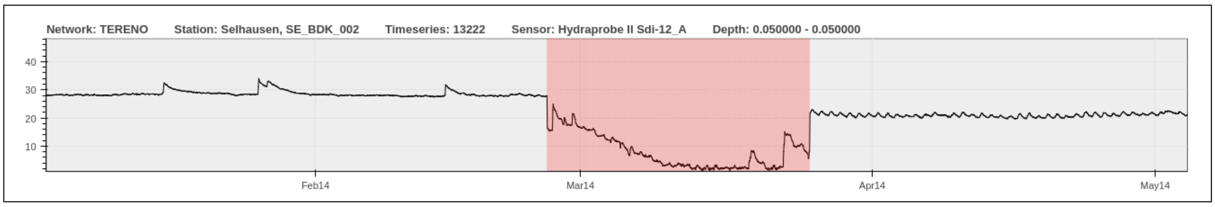


Figure 3.13: correctly flagged TERENO data set

The results (Table 3.7) for TERENO are very similar to those of SOILSCAPE.

There are many single under flagged erroneous data values and perfectly, probably manually flagged, timeseries, where suspicious observations continue over a very long time period (Figure 3.13).

Table 3.7: comparison of ISMN and TERENO flagging statistics

		flagged by ... [absolute number | % of datasets]											
variable	datasets	either		both		ismn		provider		ismn only		provider only	
sm	283,628	8,209	2.89	11	0.00	7,909	2.79	311	0.11	7,898	2.78	300	0.11
ta	49,894	6	0.01	0	0.00	0	0.00	6	0.01	0	0.00	6	0.01
ts	285,239	3,563	1.25	0	0.00	0	0.00	3,563	1.25	0	0.00	3,563	1.25
total	618,761	11,778	1.90	11	0.00	7,909	1.28	3,880	0.63	7,898	1.28	3,869	0.63

Unfortunately there also was no detailed information available.

USCRN

The station of the network U.S. Climate Reference Network (USCRN) are spread all around the United States. The USCRN network has only two number coded flags:

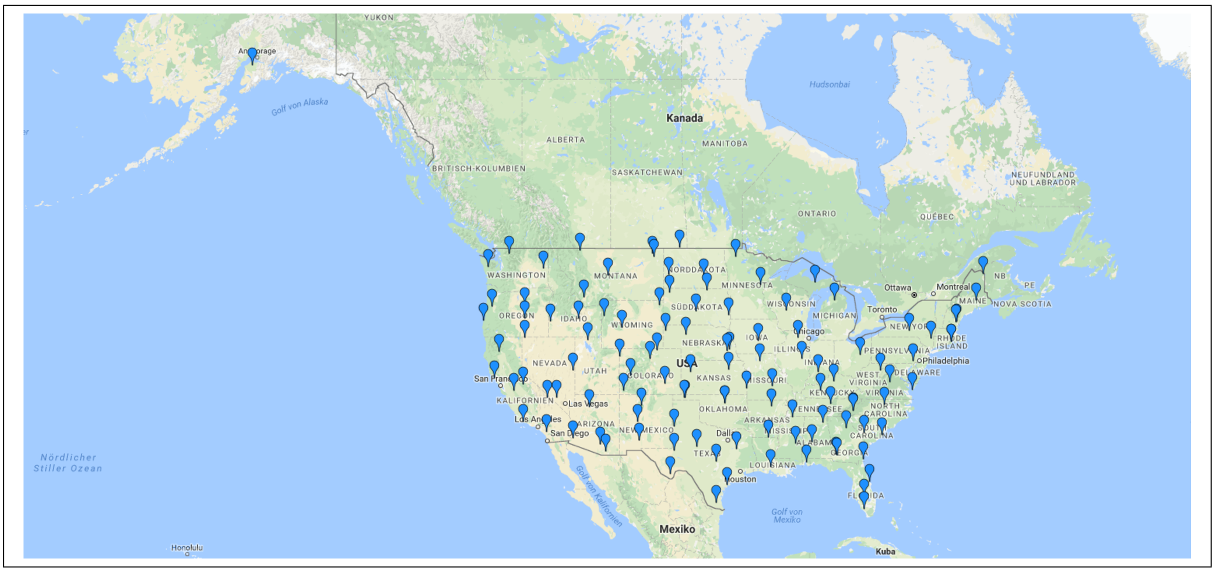


Figure 3.14: stations of the USCRN network

- 0 - good, and
- 3 - erroneous.

According to [1] there are also checks, if a value exceeds predefined boundaries. The boundaries and measurements of soil temperature by the USCRN network confirm the proper picked upper boundary of the ISMN, since the maximum value of stations located in the dessert was 57.0 °C. For soil moisture the lower boundary is 0.1 % and the upper boundary 70 %. they also check if the corresponding soil temperature value is below freezing point. Also values below 0.5 °C are flagged, since the dielectric value can be impacted of ice crystals formed around the instrument. These control method is similar to the geophysical flags of the ISMN. Also values from a known faulty sensor are flagged manually until they are replaced. Table 3.8 shows that there wasn't a single soil moisture value flagged, which contradicts their writing.

The comparison with other flagging systems was not very successful. The analyzed quality control systems are either highly over or under flagging in comparison to the ISMN. They only succeed the flagging of the ISMN, if manual flagging comes in play,

Table 3.8: comparison of ISMN and USCRN flagging statistics

		flagged by ... [absolute number | % of datasets]											
variable	datasets	either		both		ismn		provider		ismn only		provider only	
p	11,508,551	1	0.00	0	0.00	1	0.00	0	0.00	1	0.00	0	0.00
sm	24,226,741	4,561,049	18.83	0	0.00	4,561,049	18.83	0	0.00	4,561,049	18.83	0	0.00
ta	11,593,484	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
ts	26,840,301	1	0.00	0	0.00	1	0.00	0	0.00	1	0.00	0	0.00
tsf	11,634,593	24,176	0.21	1,638	0.01	12,280	0.11	13,534	0.12	10,642	0.09	11,896	0.10
total	85,803,670	4,585,227	5.34	1,638	0.00	4,573,331	5.33	13,534	0.02	4,571,693	5.33	11,896	0.01

which is simply not possible for the huge amount of data the ISMN has to offer. The positive outcome to gain from this evaluation is, that the quality control system of the ISMN is in fact very advanced and innovative.

3.4.2 Quality Control Systems from Other Soil Moisture Data Providers

The only data portal with an automated quality control system that could be found was the NASMD, a project quite similar to the ISMN, but for the United States only [9]. Their approach references to the ideas for the geophysical flags of the control system of the ISMN. For the control system of NASMD instead of GLDAS data sets, North American Land Data Assimilation System (NLDAS) (https://disc.sci.gsfc.nasa.gov/uui/datasets?keywords=NLDAS) data sets, which are only available for the unitetd states, are used. The system includes tests all soil moisture values to lie within a specific boundary and checks the corresponding porosity and soil temperature. Including the soil porosity in the quality assessment is a new approach for the ISMN for the geophysical flags, but their improvement would exceed the boundaries of this thesis. The quality control system of the NASMD does not follow a spectrum based approach.

4 New Methods

All aspects of the existing quality control methods were taken into account. The existing spectrum based flags of the ISMN were revised and six new flags were developed. Since the number of publications on quality control for soil moisture data is very sparse and therefore the ideas for new approaches were limited, a lot of effort was put in the revision of the existing spectrum based flags. For the revised and new quality flags a data viewer, with the option to show specific quality flags of the current algorithms and the new ones, was developed. Therefore the edited algorithms have a much higher impact of visual analysis for the adaption of conditions for various data sets. Six new flags could be added with ideas of the author and the ISMN data providers, who developed their own quality assessment.

4.1 Revised Algorithms

To improve the existing spectrum based algorithms boundary thresholds of most of the conditions for the error detection were loosened, because they prevented flagging of some bad data sets. Therefore other conditions were added to avoid the resulting over flagging.

4.1.1 Spikes

For the spike some conditions were added, but the initial ideas are still the foundation of this flag.

1. Since a spike is defined as a single measurement that differs significantly from the observations x_{t-1} before and x_{t+1} after x_t, the condition to also change noticeable to the next measurement was added. The thresholds of the formula concerning the value x_{t-1} were untightened, since this condition prevented many spikes from being flagged (Equation 4.1). The thresholds for the comparison of

x_t with x_{t+1} were chosen equally (Equation 4.2). A minimum of 0.5 % change in soil moisture was added to both equations to avoid an over flagging of low values.

$$\frac{x_t}{x_{t-1}} < 0.95 \quad \text{or} \quad \frac{x_t}{x_{t-1}} > 1.05 \quad \text{with} \quad x_{t-1} \neq 0 \quad \text{and} \quad |x_{t-1} - x_t| > 0.5 \tag{4.1}$$

$$\frac{x_t}{x_{t+1}} < 0.95 \quad \text{or} \quad \frac{x_t}{x_{t+1}} > 1.05 \quad \text{with} \quad x_{t+1} \neq 0 \quad \text{and} \quad |x_{t+1} - x_t| > 0.5 \tag{4.2}$$

2. In the first derivative the value before and after the spike are equal in the ideal case, but one in negative and one in positive direction. This equality is not given in real soil moisture observations and varies due to a constant rise or fall in the spectrum. Therefore it must only be about one (Equation 4.3). In the second derivative the same condition must hold. Both conditions seem redundant, but since the first derivative is rounded and takes the neighboring values into account both conditions avoid different cases of over flagging.

$$\frac{x'_{t-1}}{x'_{t+1}} \approx -1 \quad \text{with} \quad x'_{t+1} \neq 0 \tag{4.3}$$

$$\frac{x"_{t-1}}{x"_{t+1}} \approx 1 \quad \text{with} \quad x"_{t+1} \neq 0 \tag{4.4}$$

3. The third condition represents that the spikes in the second derivative are twice as big as the smaller spike at the flagged measurement. (Figure 3.2). Since a real soil moisture measurement varies a lot, the bigger spikes only have to be larger than the smaller spike.

$$\frac{x"_{t}}{x"_{t-1}} < -1 \quad \text{with} \quad x"_{t-1} \neq 0 \tag{4.5}$$

$$\frac{x"_{t}}{x"_{t+1}} < -1 \quad \text{with} \quad x"_{t+1} \neq 0 \tag{4.6}$$

4. The coefficient of variation was replaced with the mean value of the spectrum of the first derivative without the value to be flagged itself and the value before and after. It was replaced to avoid the under flagging of noisy low level soil moisture data.

$$\left|\mu([x'_{t-6}, x'_{t+6}] \backslash \{x'_{t-1}, x'_t, x'_{t+1}\})\right| < 0.5 \tag{4.7}$$

4.1.2 breaks

The ideas for the improvement of the breaks are quite similar to those of the spike. Thresholds were loosened and more conditions were added. Especially for the positive break, where over flagging is a huge problem, because of the similarity to natural soil moisture rises after precipitation events.

Negative breaks

1. The threshold of the relation of the value x_t and the value x_{t-1} was reduced to avoid under flagging. An absolute minimum difference was added to avoid over flagging of low values.

$$\frac{x_t}{x_{t-1}} < 0.9 \quad \text{with} \quad x_{t-1} \neq 0 \quad \text{and} \quad x_t - x_{t-1} < -0.5 \tag{4.8}$$

2. In the rounded first derivative the value to be flagged and the value before are equal in the ideal case. Since they vary in reality those values must be about the same for this condition. In the second derivative the same condition must hold, but x''_{t-1} must be negative. Those conditions would be equal if the first derivation was not rounded. Because it is rounded both conditions complete one another, because the first derivative takes its neighboring values into account and the second derivative does not.

$$\frac{x'_t}{x'_{t-1}} \approx 1 \quad \text{with} \quad x'_{t-1} \neq 0 \tag{4.9}$$

$$\frac{x''_t}{x''_{t-1}} \approx -1 \quad \text{with} \quad x''_{t-1} \neq 0 \tag{4.10}$$

3. The first formula of the third condition was already used for the existing flagging method (Equation 4.11). Since the same condition must hold after the break it was also added (Equation 4.11). The division was inverted, since x''_{t-1} and x''_t should not be zero anyway and x''_{t-2} and x''_{t+1} are often zero. Again, the

thresholds were loosened.

$$\left|\frac{x''_{t-2}}{x''_{t-1}}\right| < 0.15 \quad \text{with} \quad x''_{t-1} \neq 0 \tag{4.11}$$

$$\left|\frac{x''_{t+1}}{x''_{t}}\right| < 0.15 \quad \text{with} \quad x''_{t} \neq 0 \tag{4.12}$$

4. The spectrum before and after the break must be rather smooth to avoid over flagging. This condition replaces the coefficient of the variance since the boundary was too big for high level soil moisture values and too small for low level values. Furthermore the first derivative serves a good purpose, because it is rounded and a single outlier is smoothed.

$$\left|\mu([x'_{t-6}, x'_{t+6}] \backslash \{x'_{t-1}, x'_{t}\})\right| < 0.5 \tag{4.13}$$

5. The last equation only differs from the existing methods by adding the value before x'_t, to consider that the first derivative is rounded.

$$x'_t + x'_{t-1} < -10 \cdot \left|\mu([x'_{t-6}, x'_{t+6}] \backslash \{x'_{t}, x'_{t-1}\})\right| \tag{4.14}$$

Positive Breaks

The positive break does not differ much from the negative break. One condition was added since the positive break is much more difficult to detect, due to its similarity to a natural rise caused by a precipitation event. The other formulas are similar, but some are altered for the positive direction of the break and are listed for completion.

1. For the positive break the relation between x_t and x_{t-1} must be higher than a certain threshold. Also the same absolute difference as for the negative break must hold.

$$\frac{x_t}{x_{t-1}} > 1.1 \quad \text{with} \quad x_{t-1} \neq 0 \quad \text{and} \quad x_t - x_{t-1} > 0.5 \tag{4.15}$$

2. The second condition is exactly the same as for the negative break.

$$\frac{x'_t}{x'_{t-1}} \approx 1 \quad \text{with} \quad x'_{t-1} \neq 0 \tag{4.16}$$

$$\frac{x"_t}{x"_{t-1}} \approx -1 \quad \text{with} \quad x''_{t-1} \neq 0 \tag{4.17}$$

3. The third condition is also the same. Again, it is essential to put x''_{t-1} and x''_t as the denominator, since x''_{t-2} and x''_{t+1} are often zero.

$$\left|\frac{x"_{t-2}}{x"_{t-1}}\right| < 0.15 \quad \text{with} \quad x"_{t-1} \neq 0 \tag{4.18}$$

$$\left|\frac{x"_{t+1}}{x"_t}\right| < 0.15 \quad \text{with} \quad x"_t \neq 0 \tag{4.19}$$

4. The spectrum around the break must also be very smooth.

$$\left|\mu([x'_{t-6}, x'_{t+6}] \backslash \{x'_{t-1}, x'_t\})\right| < 0.5 \tag{4.20}$$

5. For the positive break the addition of x'_t and x'_{t-1} must be ten times bigger than the spectrum of the first derivative.

$$x'_t + x'_{t-1} > 10 \cdot \left|\mu([x'_{t-6}, x'_{t+6}] \backslash \{x'_t, x'_{t-1}\})\right| \tag{4.21}$$

6. The last condition was added to avoid the over flagging of natural precipitation events or snow melting processes. After a break soil moisture stays at a rather constant level. Due to an ongoing natural event the soil moisture often either rises further, because of a continuing rain fall, or drops due to drying processes. These situations are mostly excluded by the following formulas.

$$\frac{x_t}{x_{t+1}} > 0.99 \qquad \text{and} \qquad \frac{x_t}{x_{t+1}} < 1.01 \qquad \text{with} \qquad x_{t+1} \neq 0 \tag{4.22}$$

$$\frac{x_{t+1}}{x_{t+2}} > 0.99 \qquad \text{and} \qquad \frac{x_{t+1}}{x_{t+2}} < 1.01 \qquad \text{with} \qquad x_{t+2} \neq 0 \tag{4.23}$$

4.1.3 Plateaus

Low Level Plateaus

The naming of this flag was changed from "low constant values" to "low level plateau" to avoid a confusion with a new flag called "constant values", which will be described in section 4.2.

1. The first condition for a low level plateau is still a negative break as the plateau start x_t with $t = t_{pl\_start}$.

2. The coefficient of variation was replaced with the variance itself, since low level plateaus mostly occur at a low level of soil moisture observations. The coefficient of variation is a relation that allows a higher variation at higher data reading. With the variance itself more low level values and less high level values will be flagged. The boundary was also loosed, because the over flagging of low level plateaus was never a problem, but the under flagging was.

$$\sigma^2([x_t, x_{t+n}]) < 0.01 \quad \text{with} \quad n \geqslant 12 \quad \text{and} \quad t = t_{pl\_start} \tag{4.24}$$

3. A third condition was added to assure a soil moisture rise at the end of the plateau to avoid the over flagging of natural drying process that are rather slow.

$$x'_t > 0 \quad \text{with} \quad t = t_{pl\_end} \tag{4.25}$$

Saturated Plateaus

Saturated Plateaus are very easily mistaken for natural slow drying process. The conditions were much adapted with the result of a much higher detection rate, but also an increasing but acceptable over flagging.

1. As before a soil moisture rise must have occurred for a saturated plateau (Equation 3.12), but in the new method it has to happen in the previous three hours and the rise has to be higher than for the current quality flags. It is varying when the actual plateau starts after a heavy rain fall, but visual inspection of

many timeseries showed that it does not lie back more than three hours.

$$\sum_{i=0}^{3}(x_{t-i-1} - x_{t-i}) > 0.5 \quad \text{and} \quad x_t - x_{t-1} > 0.1 \quad \text{with} \quad t = t_{pl\_start} \tag{4.26}$$

2. The variance of the plateau is now considered relative with the coefficient of variation, because low soil moisture readings were highly over flagged and very high saturated plateaus, which show a high variation could not be detected. The variance of the whole plateau and every twelve hour part are calculated separately. The additional boundary variance for the whole plateau avoids over flagging of drying process, but if the plateau lasts over a long time period an appended value would not have much influence. Therefore both boundaries are valuable assets.

$$\frac{\sigma^2([x_t, x_{t+n}])}{\mu([x_t, x_{t+n}])} < 0.0075 \quad \text{with} \quad n \geqslant 12 \quad \text{and} \quad t = t_{pl\_start} \tag{4.27}$$

$$\frac{\sigma^2([x_{t-12}, x_t])}{\mu([x_{t-12}, x_t])} < 0.005 \quad \forall\{t \,|\, t_{pl\_start} + 12 \leqslant t \leqslant n\} \tag{4.28}$$

3. To detect high soil moisture differences of two neighboring values, which should not occur within a plateau, a new condition was added. The relative difference also helps to avoid under flagging of high soil moisture plateaus and over flagging of correct low level data.

$$\frac{x_i - x_{i-1}}{\mu([x_t, x_{t+n}])} < 0.01 \quad \forall x_i \in [x_{t+1}, x_{t+n}] \quad \text{with} \quad n \geqslant 12 \quad \text{and} \quad t = t_{pl\_start} \tag{4.29}$$

4. The last condition is still the same as for the existing methods. Plateaus are only flagged as such if they are at the upper five percent of the soil moisture readings of a timeseries

$$\mu([x_{pl\_start}, x_{pl\_end}]) > \max([x_{t_0}, x_{t_{end}}] \backslash \{x_t > 100\}) \cdot 0.95 \tag{4.30}$$

4.2 New Algorithms

The new algorithms are based on observations during the revision of the existing flags or ideas of network providers for their on quality control system. They are a valuable addition to avoid undetected erroneous data sets (Table 4.1).

Table 4.1: additional spectrum based ISMN quality flags

variable	flag	condition
soil moisture	D11	suspicious value before NaNs
	D12	suspicious value after NaNs
	D13	severe soil moisture drop
	D14	alternating values
	D15	constant values
	D16	highly flagged spectrum

4.2.1 Suspicious Values Around Missing Values

Values before or after a sensor failure are often outliers. Especially after long time periods of a sensor drop out.

Suspicious Values Before Missing Values

The first and second derivatives are quite similar to those of breaks (Figure 4.1).

1. A sensor failure for a few measurements normally does not lead to suspicious values before or after this occurrence. To avoid over flagging a minimum of a 3 hour sensor drop out is necessary.

$$\nexists\,[x_{t+1}, x_{t+n}] \quad \text{with} \quad n \geqslant 3 \tag{4.31}$$

2. The difference to the value before the suspicious value must be higher than a certain relative boundary, which was defined to be more than 5 %.

$$\frac{x_t}{x_{t-1}} < 0.95 \quad \text{or} \quad \frac{x_t}{x_{t-1}} > 1.05 \quad \text{with} \quad |x_t - x_{t-1}| > 0.5 \quad \text{and} \quad x_{t-1} \neq 0 \tag{4.32}$$

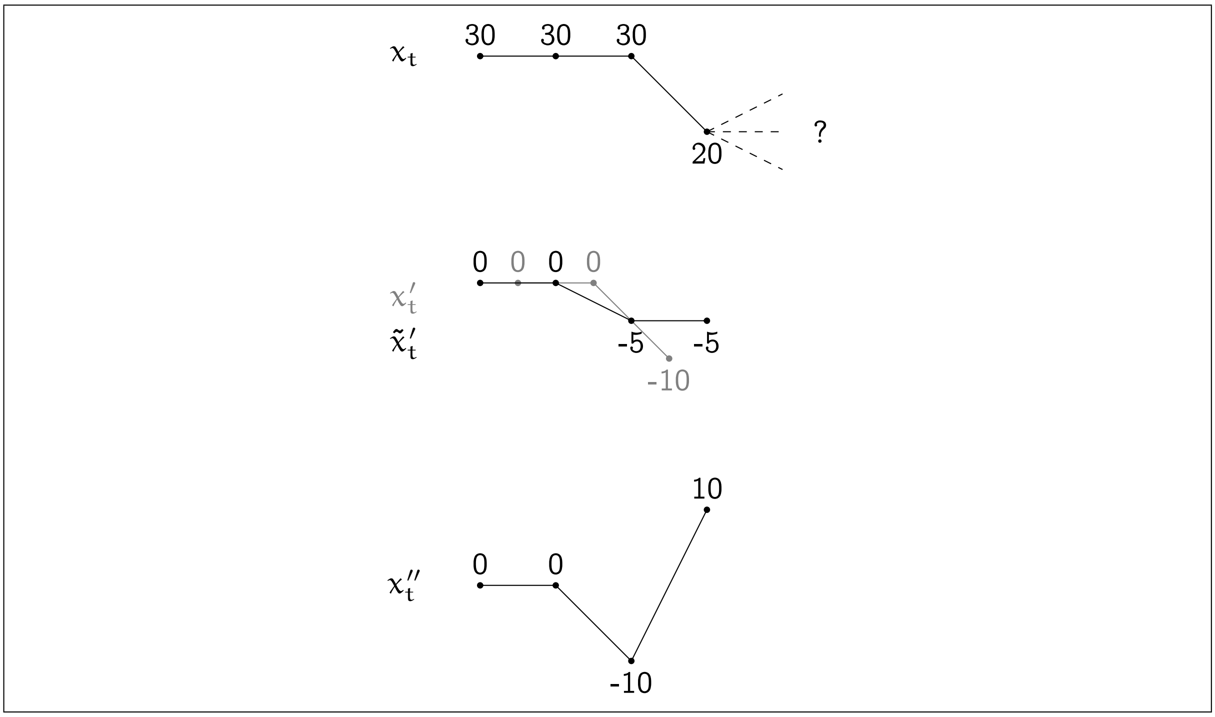


Figure 4.1: data set x_t with a suspicious value before missing values and its first derivative $\tilde{x}'_t$ and second derivative x''_t

3. The spectrum before the suspicious value has to be smooth to differ between natural soil moisture varieties and dubious values.

$$x'_{t-1} < -10 \cdot \left|\mu([x'_{t-2}, x'_{t-6}])\right| \qquad \text{with} \qquad \left|\mu([x'_{t-2}, x'_{t-6}])\right| < 1 \tag{4.33}$$

4. As for a break, the first derivative at the suspicious value must be much higher than the past spectrum.

$$x'_{t-1} + x'_t > 10 \cdot \left|\mu([x'_{t-2}, x'_{t-6}])\right| \qquad \text{with} \qquad \left|\mu([x'_{t-2}, x'_{t-6}])\right| < 1 \tag{4.34}$$

Suspicious Values After Missing Values

The algorithms for suspicious values after NaN-Values (Figure 4.2) are analog to those for suspicious values before NaN-Values.

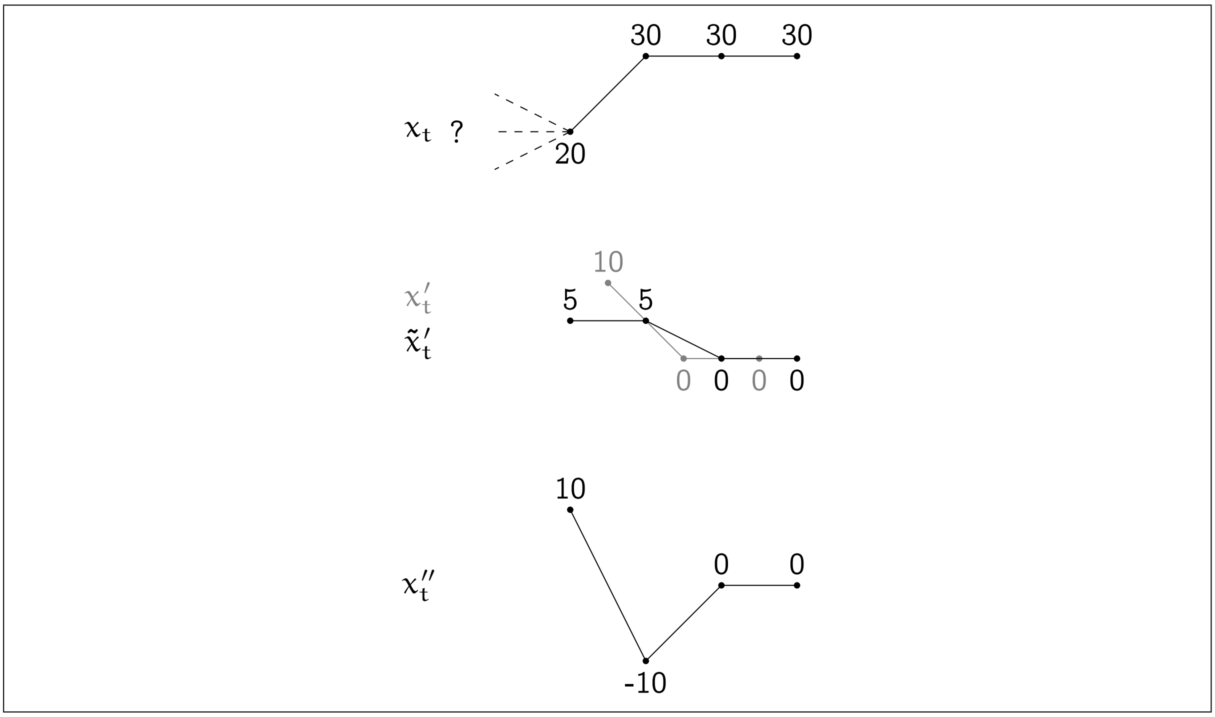


Figure 4.2: data set x_t with a suspicious value after missing values and its first derivative $\tilde{x}'_t$ and second derivative x''_t

4.2.2 Severe Soil Moisture Drop

Due to the slow soil moisture drying process it seems apparent to introduce a flag that triggers, if a negative soil moisture change within one hour is beyond a certain threshold. Compared to a negative break there are no other conditions to be full filled, but the drop must be much higher. To avoid over flagging of high soil moisture values and under flagging of low values, a relative boundary was set that muss be succeeded. It was set to be more than 25 %. A minimum absolute boundary was defined to avoid the under flagging of low soil moisture values.

$$\frac{x_t}{x_{t-1}} < 0.75 \quad \text{with} \quad x_{t-1} \neq 0 \quad \text{and} \quad x_t - x_{t-1} < -0.5 \tag{4.35}$$

4.2.3 Alternating Values

In some soil moisture timeseries a sensor malfunction occurs that results in alternating values. The readings constantly switch between rather high and rather low soil moisture readings. Five conditions have to be met to flag such alternating values.

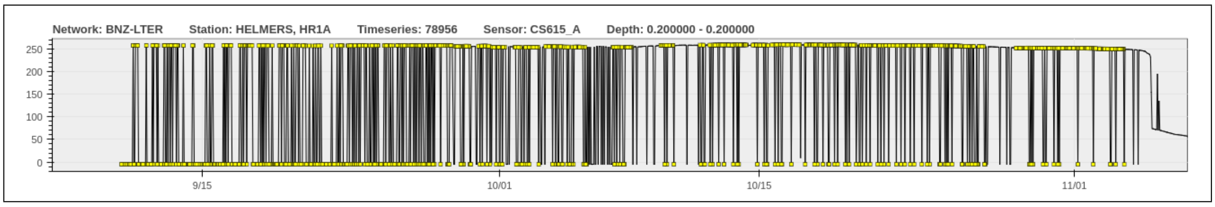


Figure 4.3: alternating values

1. The upper and lower soil moisture readings are parted in two sets, which both have to include at least three measurements their union is the whole erroneous data period with at least thirteen values.

$$\exists\{X\} : |X| \geqslant 3 \quad \text{and} \quad \exists\{Y\} : |Y| \geqslant 3 \quad \text{and} \quad X \cup Y = [x_t, x_{t+n}] \quad \text{with} \quad n \geqslant 12 \tag{4.36}$$

2. Each group must include a value that has a predecessor and a successor in time of the other group.

$$\exists x \in X : t_{y1} < t_x < t_{y2} \qquad y1, y2 \in Y \tag{4.37}$$

$$\exists y \in Y : t_{x1} < t_y < t_{x2} \qquad x1, x2 \in X \tag{4.38}$$

3. The variance of each group must be less than 0.5%.

$$\sigma^2(X) < 0.5 \qquad \wedge \qquad \sigma^2(Y) < 0.5 \tag{4.39}$$

4. The relative difference between the sets must be at least 25%.

$$\frac{\mu(X)}{\mu(Y)} < 0.75 \qquad \vee \qquad \frac{\mu(X)}{\mu(Y)} < 1.25 \tag{4.40}$$

5. Each value of each set must differ from the mean of the set by at most 1 % to avoid outliers that should not occur in such a set and therefore avoid over flagging.

$$|\mu(X) - x| < 1 \qquad \forall x \in X \tag{4.41}$$

$$|\mu(Y) - y| < 1 \qquad \forall y \in Y \tag{4.42}$$

4.2.4 Constant Values

Soil moisture values are highly variable. Even if there occurs no precipitation event, they are not only influenced by a yearly cycle, but also by daily radiation and temperature variations. Therefore it is very unlikely for soil moisture readings to remain constant for more than 24 hours.

The only condition for constant values is that the measured value does not change by any digit over at least three days, where three daily cycles should be visible.

$$\forall x \in [x_t, x_{t+n-1}] : x = x_{+1} \qquad \text{with} \qquad n \geqslant 72 \tag{4.43}$$

4.2.5 Highly Flagged Spectrum

There were many random erroneous observations detected that can not be detected by specified algorithms. This often concerns data values that lie within a spectrum were already most of the data sets are flagged. Therefore a flag is introduced that triggers, if more than half of the data sets within 24 hours before and after are already detected as suspicious values.

5 Evaluation

For the evaluation of the revised and new algorithms for the quality control system many test data sets were chosen and manually flagged to be compared against the existing and the new methods. Many examples will be discussed why the improved algorithms result in a better flagging statistic and why some methods fail at specific data sets.

5.1 Results

From each network one timeseries of one year was chosen as a test data set, to assure various sensors, depths and climate and land cover classes. Ten networks AACES, AWDN, CHINA, ICN, IOWA, KHOREZM, MONGOLIA, PBO_H2O, RUSWET-AGRO, RUSWET-GRASS and RUSWET-VALDAI, were excluded since their observations are not hourly. Therefore 44 data sets were used for the validation (Table 5.1). The chosen data sets were not used for developing the algorithms. Every test data set was manually flagged and afterwards compared to the existing and the new quality flags by calculating an error matrix for both, the existing and the new flags respectively.

The error matrix for the existing quality flags shows that many erroneous data sets were not detected as such (Table 5.2). The error matrix of the edited and new methods shows a much higher detection rate (Table 5.3).

Table 5.1: test data sets for the evaluation of the existing and new ISMN quality flags

network	station	sensor	depth [m]	year	error types
AMMA-CATCH	Belefoungou-Mid	CS616	0.40 - 0.40	2010	D06 D12
ARM	Hillsboro	Water Matric Potential Sensor	0.60 - 0.60	1997	D06 D15
BIEBRZA_S-1	grassland_soil_2	GS-3	0.05 - 0.05	2016	D10
BNZ-LTER	HELMERS	CS615	0.20 - 0.20	2001	D06 D10 D13 D14 D15
CALABRIA	Chiaravalle Centrale	ThetaProbe ML2X	0.30 - 0.30	2008	D06 D07 D09 D15
CAMPANIA	Bagnoli	ThetaProbe ML2X	0.30 - 0.30	2011	D06 D11 D12 D13
CARBOAFRICA	SD-DEM	CS615	0.30 - 0.30	2003	D07 D08 D09
COSMOS	Shale Hills	Cosmic-ray Probe	0.00 - 0.16	2016	D06 D07
CTP_SMTMN	L01	EC-TM	0.00 - 0.05	2011	
DAHRA	DAHRA	ThetaProbe ML2X	0.10 - 0.10	2005	
FLUXNET-AMERIFLUX	Tonzi Ranch	ThetaProbe ML2X	0.20 - 0.20	2012	D08 D10
FMI	SOD101	5TE	0.05 - 0.05	2016	D10
GTK	Porill	CS616	0.70 - 0.70	2009	D07 D10
HOBE	1.07	Decagon 5TE	0.50 - 0.55	2010	D10
HSC_SELMACHEON	Selmacheon	HydraProbe Analog	0.00 - 0.10	2008	D06 D07 D08 D10 D13
HYDROL-NET_PERUGIA	WEEF 2	TRASE-BE	0.25 - 0.25	2013	D06 D07 D08 D14
HYU_CHEONGMICHEON	Suresan	HydraProbe T100A	0.00 - 0.10	2011	D13
IIT_KANPUR	ITTK_Airstrip	Water Scout SM100	0.50 - 0.50	2012	D07 D08 D12
iRON	Brush Creek	10HS	0.20 - 0.20	2015	D07 D09
LAB-net	Oromo Calibration Site	Sensor CS650	0.07 - 0.07	2014	D15
MAQU	NST_05	ECH20 EC-TM	0.05 - 0.05	2009	D11
METEROBS	Monte Pino	EnvironSCAN	0.30 - 0.30	2011	
MOL-RAO	Falkenberg	TRIME-EZ	0.60 - 0.60	2008	D06
ORACLE	SUIZY	TRASE 16	0.70 - 0.70	2011	D07 D09
OZNET	Alabama	HydraProbe	0.00 - 0.05	2010	D11 D13
REMEDHUS	El Tomillar	HydraProbe	0.00 - 0.05	2015	D07 D13
RSMN	Banloc	5TE	0.00 - 0.05	2016	D06 D11 D13
SASMAS	Brunbrae	HydraProbe	0.00 - 0.30	2007	
SCAN	Kemolche Gulch	HydraProbe Analog	0.10 - 0.10	2012	D06 D07 D08 D13
SKKU	SKKU_Jinwicheon_6	5TM	0.10 - 0.10	2014	D06 D07 D08 D10 D13
SMOSMANIA	Carbieres dAvignon	ThetaProbe ML2X	0.20 - 0.20	2015	D06 D07 D08
SNOTEL	WHISKEY CK	HydraProbe Analog	0.50 - 0.50	2016	D06 D13
SOILSCAPE	node1019	EC5	0.42 - 0.42	2014	D12 D10
SWEX_POLAND	P1	D-LOG-mpts	0.10 - 0.10	2012	D07 D08 D13 D14
SW-WHU	BoaxieBDS-RSoilMoisture	SoilMoisture-AverageSensor	0.10 - 0.10	2015	D06 D11
TERENO	Selhausen	Hydraprobe II SDI-12	0.20 - 0.20	2014	D12
UDC_SMOS	Erlbach	IMKO TDR	0.05 - 0.05	2010	D06 D07
UMBRIA	Monterchi	EnviroSCAN	0.15 - 0.25	2011	D11
UMSUOL	San Pietro Capofiume	TDR 100	0.45 - 0.45	2009	D06 D07 D08 D15
USCRN	Dinosaur_2_E	HydraProbe II SDI-12	0.10 - 0.10	2011	D06 D07 D08 D10
USDA-ARS	Little River	HydraProbe Analog	0.00 - 0.05	2003	
VAS	MelbexI	HydraProbe	0.00 - 0.05	2010	D11
WEGENERNET	15	pF-Meter	0.30 - 0.30	2016	D06 D10
WSMN	WSMN_4	CS615	0.03 - 0.03	2015	D06 D07 D08 D09

Table 5.2: error matrix of 244502 data values for the existing quality flags

	flagged ... [absolute number | % of datasets]							
flag	manual		manual		¬ manual		¬ manual	
	automatic		¬ automatic		automatic		¬ automatic	
D06	16	11.0	129	89.0	215	0.1	243597	99.9
D07	35	31.0	78	69.0	14	0.0	243798	100.0
D08	23	27.1	61	71.8	44	0.0	243768	100.0
D09	192	5.0	3342	86.9	845	0.4	242967	99.6
D10	1217	24.1	3823	75.9	398	0.2	243414	99.8
any	1563	9.1	15701	90.9	1435	0.6	225113	99.4

Table 5.3: error matrix of 244502 data values for the improved quality flags

	flagged ... [absolute number | % of datasets]							
flag	manual		manual		¬ manual		¬ manual	
	automatic		¬ automatic		automatic		¬ automatic	
D06	114	78.6	31	21.4	72	0.0	244430	100.0
D07	63	55.8	50	44.2	19	0.0	244483	100.0
D08	42	49.4	42	49.4	22	0.0	244480	100.0
D09	3020	89.8	344	10.2	48	0.0	244454	100.0
D10	2579	51.2	2461	48.8	1993	0.8	242509	99.2
D11	5	50.0	5	50.0	2	0.0	244500	100.0
D12	16	76.2	5	23.8	0	0.0	244502	100.0
D13	284	94.7	15	5.0	141	0.1	244361	99.9
D14	1143	79.6	293	20.4	1	0.0	244501	100.0
D15	10253	100.0	0	0.0	0	0.0	244502	100.0
any	14326	83.8	2768	16.2	1861	0.8	225547	99.2

5.2 Discussion

Th under flagging of the existing methods can be explained with the fact that the algorithms were developed when there were not so many data sets integrated in the ISMN and the huge variety of sensors, depths, land cover and climate classes was not given. The revised methods show a higher detection rate and also the over flagging could be reduced in most cases. Only the saturated plateaus show a higher over flagging, but that can be justified with the high detection rate of additional over a thousand flagged data values. For a validation of satellite data thousands of under flagged erroneous data values would be much more severe than missing good values.

5.2.1 Spikes

Many spikes were not flagged with the existing methods, because the boundaries for the relation between the value of the spike and the value before was defined very tight (Equation 3.1). In Figure 5.1 the existing flags are compared to the revised methods in the plot below.

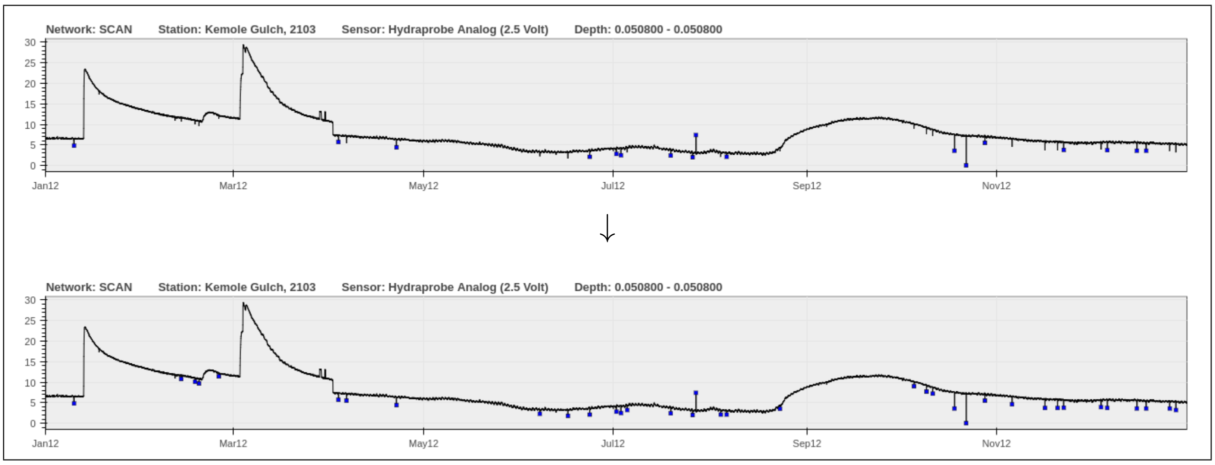


Figure 5.1: previous under flagged spikes

If only the boundary of the relation would have been loosened many data sets would have been over flagged. With more conditions this could be avoided and even values that were over flagged before could be saved from the incorrect assessment. One important addition is the consideration of the high difference between the spike itself and the value after, which was not considered for the existing method (Figure 5.2).

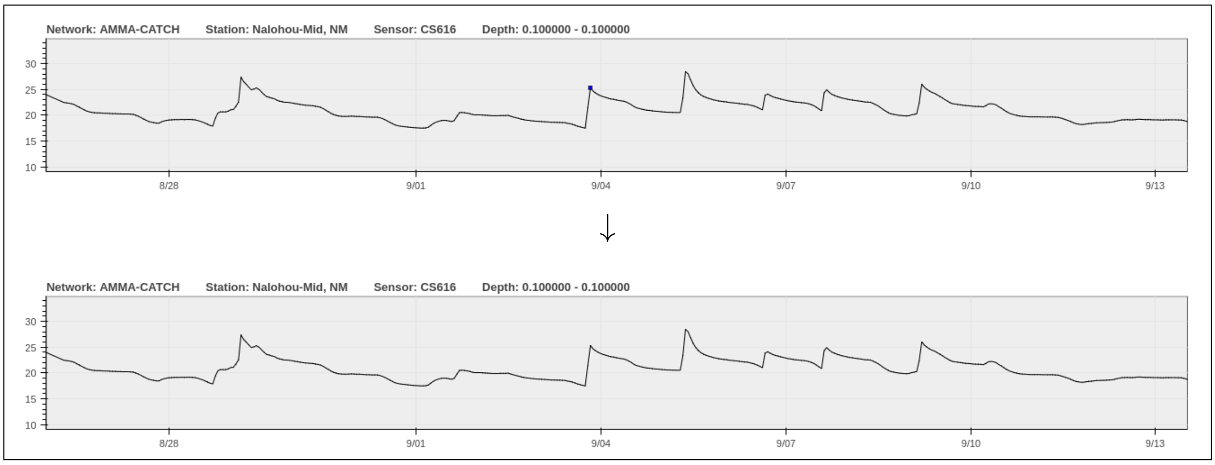


Figure 5.2: previous over flagged spike

If there are two spikes in a row they cannot be flagged due to the revised conditions for the first and second derivative. Figure 5.3 shows a spike that was flagged and a spike before and another spike after that both look very alike, but cannot be flagged, because they follow another smaller spike.

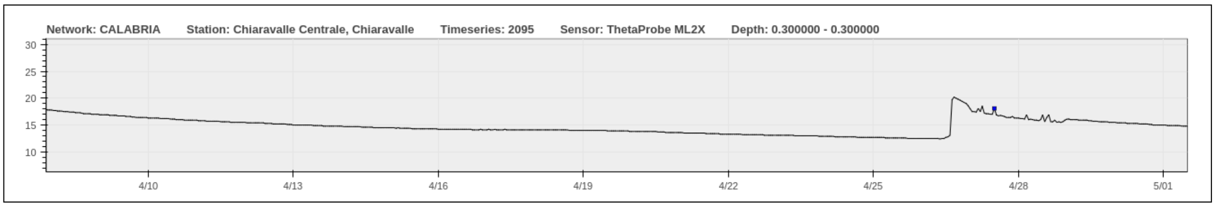


Figure 5.3: example for a spike that is still not flagged

Most of the possibly over flagged spikes of the error matrices of both methods result from dubious data sets, where the perception of whether a value is a spike or not was very subjective. In Figure 5.4 for example, the sensor readings are only at a full percentage. Therefore what seems to be a spike is probably only a rounded value. In all the example timeseries was no obvious correct data set found that was over flagged.

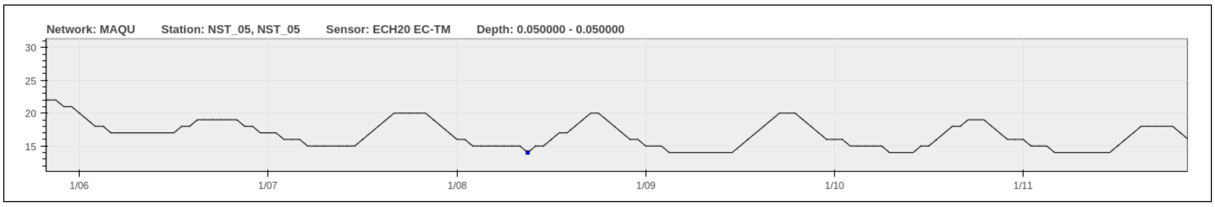


Figure 5.4: example for a correct value that is still flagged as a spike

5.2.2 Breaks

Negative Breaks

Many negative breaks were not flagged with the existing methods, because the conditions for the second derivative included a division with a value in the denominator that is most likely to be zero, especially in ideal cases (Figure 5.5). This problem was fixed by inverting the division and the corresponding boundary for the revised methods.

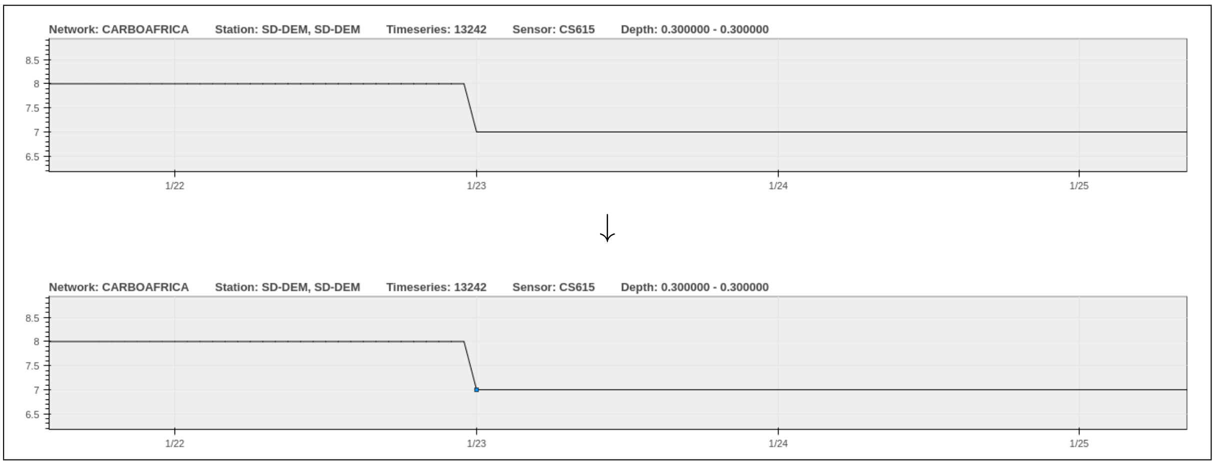


Figure 5.5: previous under flagged negative break

There are some negative breaks that still cannot be detected with both methods, because the break happens over two steps (Figure 5.6). The first and second derivative are altered and the conditions do not hold. Also the variance of the spectrum before the break is often not small enough, because of this intermediate step.

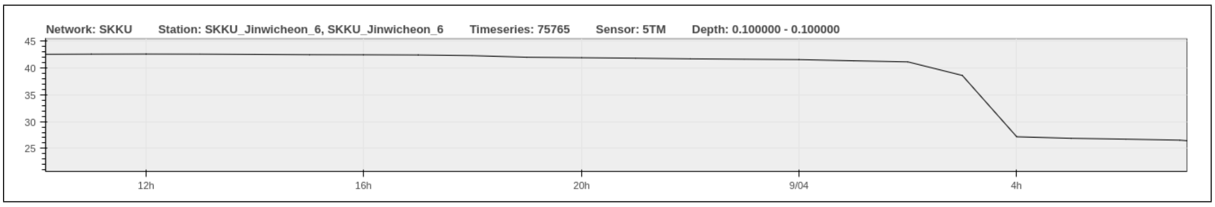


Figure 5.6: example for a negative break that is still not flagged

The over flagging of negative breaks was not a problem with the existing methods. It was a bit increased with the edited methods to assure the flagging of erroneous data. This especially applies for noisy data, where the determination of erroneous data is very subjective (Figure 5.7).

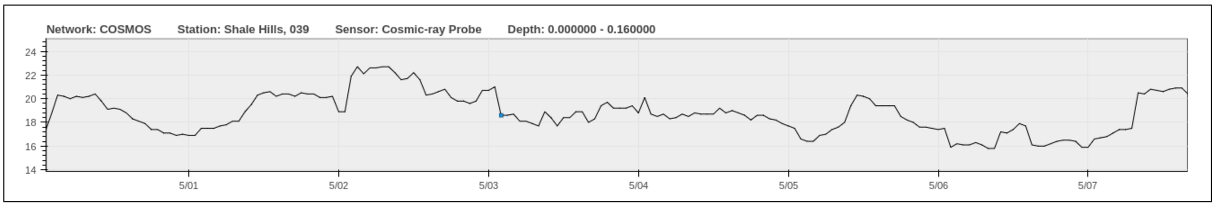


Figure 5.7: example for a correct value that is still flagged as a negative break

The flagging of negative breaks could also be improved. Nearly twice as much breaks are now detected. The over flagging of suspicious noisy data, which is suspicious anyway, was accepted to increase the flagging of erroneous data.

Positive Breaks

Also positive breaks were under flagged with the existing methods, because the division by zero was not considered. That the first derivative between the discrete points is rounded had no influence in the conditions as well. When the variance of the first derivative around the value to be flagged was tested to be much smaller than at x_t the value x_{t-1} was included (Equation 3.5, which was also high due to the rounding of the first derivative. Therefore often high breaks were not marked as such, because the first derivative before the value to be flagged was too high for the variance to be small enough. With the improved conditions more positive breaks are now flagged (Figure 5.8).

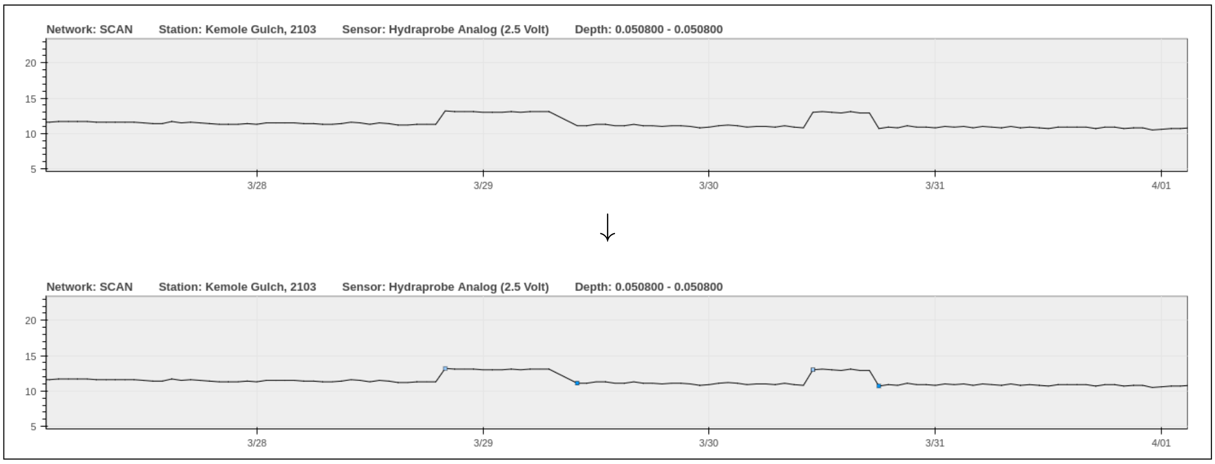


Figure 5.8: previous under flagged positive breaks

A huge problem with positive breaks is the over flagging for both methods. Precipitation events or snow melting processes cause a sever soil moisture rise, which is hard to separate from sudden positive breaks with automated algorithms. A condition was added that the difference between the two time steps after the break has to be very small to reduce the over flagging (Figure 5.9).

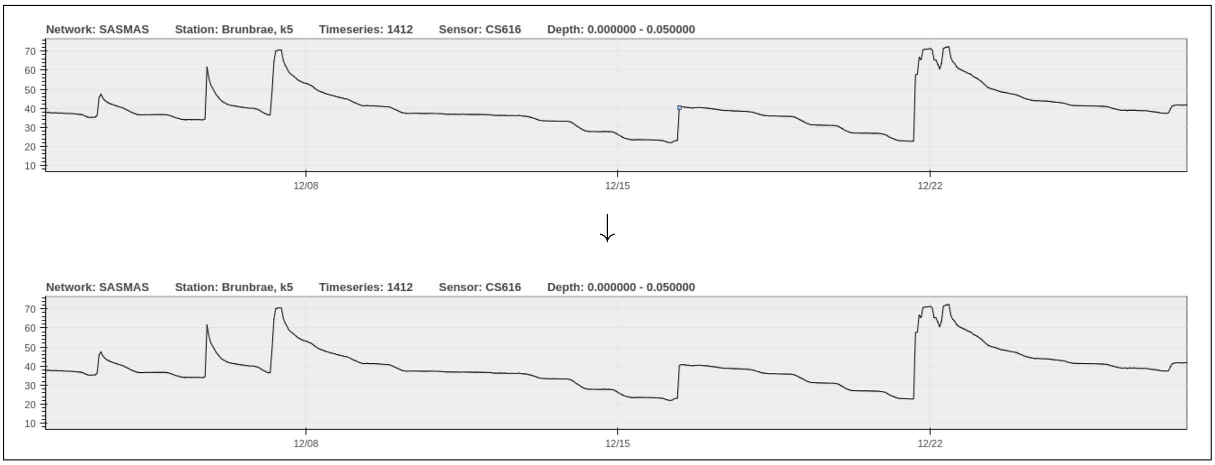


Figure 5.9: previous over flagged positive break

With both methods breaks are often not flagged if they occur in a constant rising spectrum (Figure 5.10). The condition for a small variance does not hold, which is very important to separate natural events from breaks, especially for positive breaks.

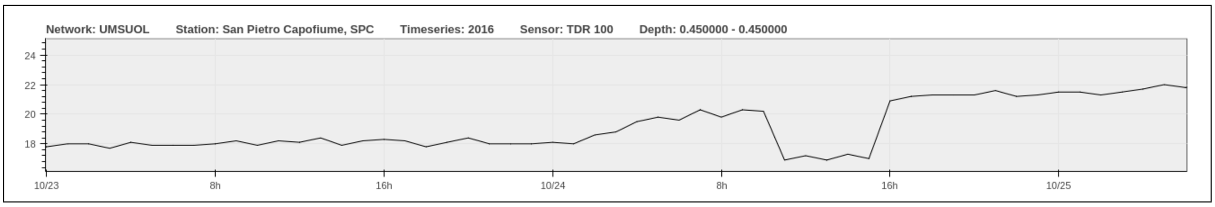


Figure 5.10: example for a positive break that is still not flagged

The over flagging of positive break is still a problem. If the drying process of soil is very slow or the rain continues to hold soil moisture at a constant level, the difference between those natural events and a positive break cannot be detected with automated methods (Figure 5.11). The only obvious solution would be a manual control of those flags, but the time and effort for a data portal as big as the ISMN cannot be justified.

While negative breaks are easily detected, it is much harder to separate positive breaks from natural precipitation events. The detection rate of negative breaks was already very good, but still could be improved. The flagging of positive breaks is still not optimal, but much better than with the existing methods.

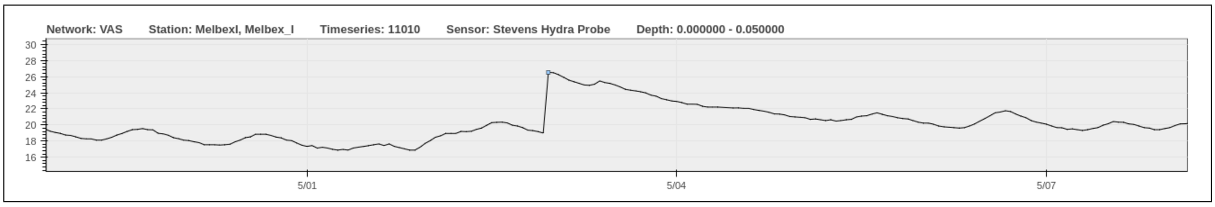


Figure 5.11: example for a correct value that is still flagged as a positive break

5.2.3 Plateaus

Low Level Plateaus

The under flagging of low level plateaus could be reduced by increasing the allowed variance over the whole plateau. Also the value for the variance was chosen absolute instead of relative since low level plateaus occur at low soil moisture readings and a value relative to the level of soil moisture readings would lead to over flagging at high readings and under flagging at low values. Figure 5.12 shows an example of a plateau that was not fully flagged with the existing methods.

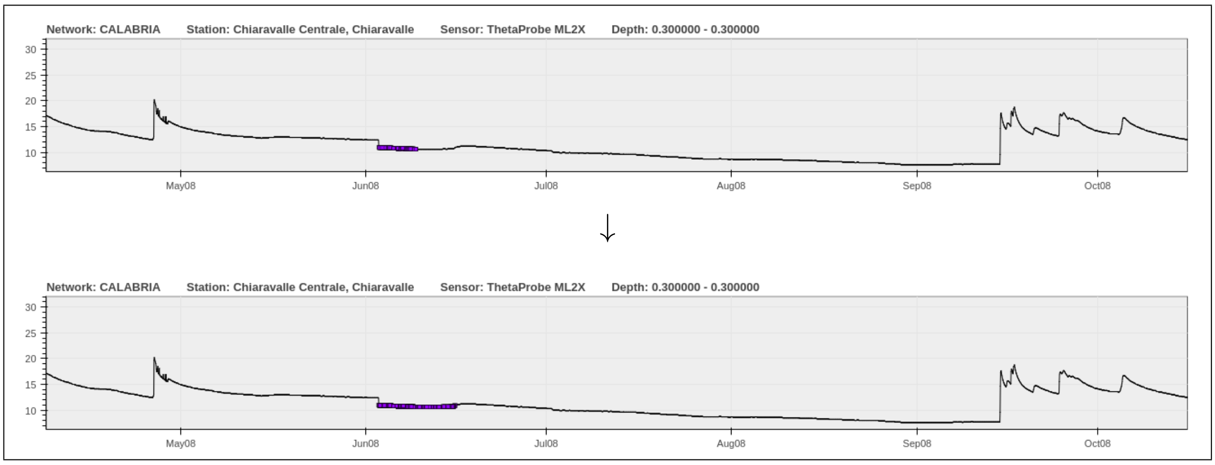


Figure 5.12: previous under flagged low level plateau

Because of this relative variance high level soil moisture observations were over flagged with the existing methods (Figure 5.13). Another reason was the missing condition for the plateau to end with a rise of soil moisture. With this condition at least the over flagging of the last plateau could have been avoided.

Very noisy data still cannot be flagged without highly over flagging correct data, because the boundary for the variance would have to be chosen much to high (Figure 5.14).

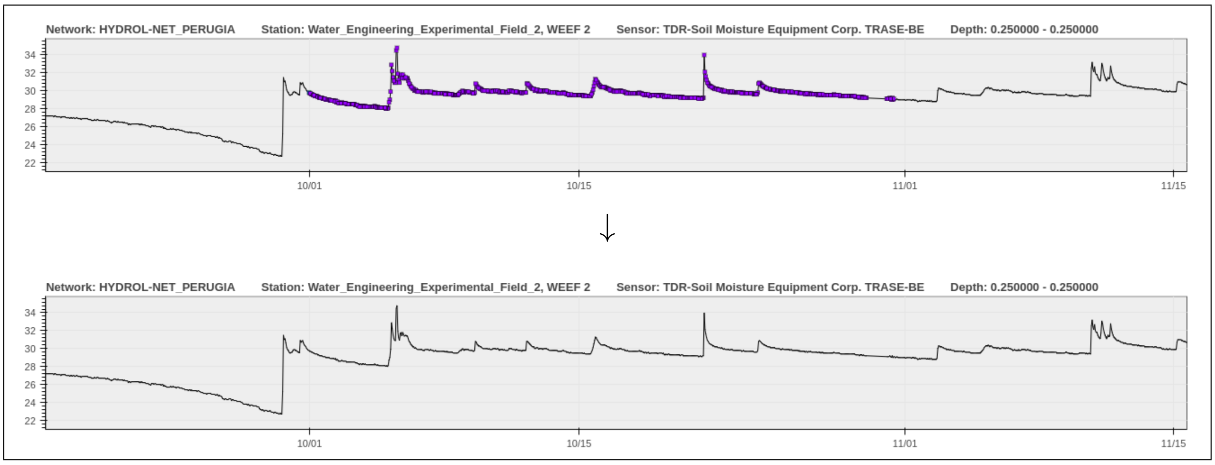


Figure 5.13: previous over flagged low level plateaus

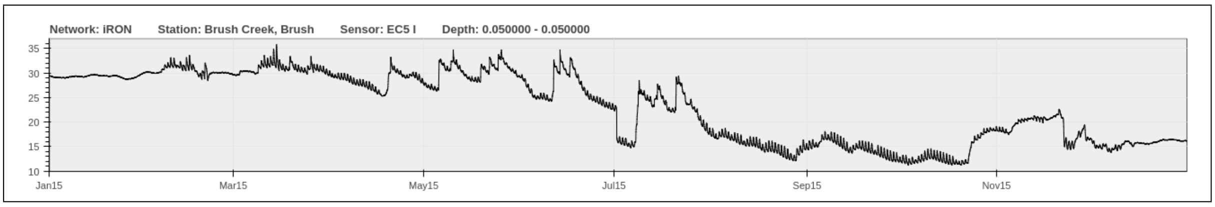


Figure 5.14: example for a low level plateau that is still not flagged

If a timeseries includes alternating positive and negative breaks the incorrect data level can hardly be detected with algorithms. In Figure 3.5 the lower level is most likely the correct measurement and the two series at a higher level the erroneous ones.

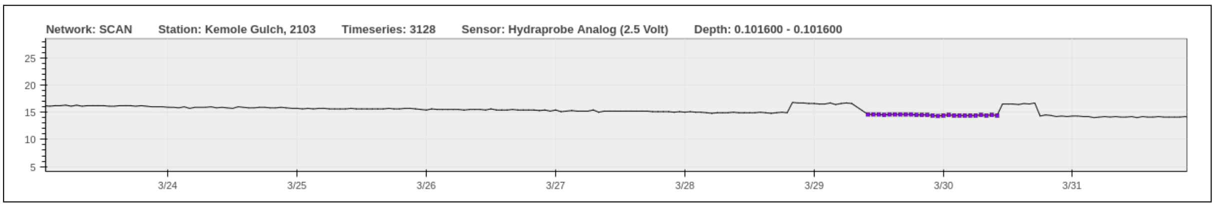


Figure 5.15: example for correct values that are still flagged as low level plateau

Improving the low level plateau flag was very successful. The percentage of detected erroneous data could be raised from 5 % to 76 %, while simultaneously even the over flagging could be reduced.

Saturated Plateaus

Saturated plateaus are the most difficult suspicious series of measurement to detect. An ideal saturated plateau is nearly constant, but in real measurements they actually have a very high variation. For the revised flags the variation boundary was defined relative to the level of soil moisture, since there are higher variations at higher soil moisture observations. With this method many more saturated plateaus can be detected (Figure 5.16).

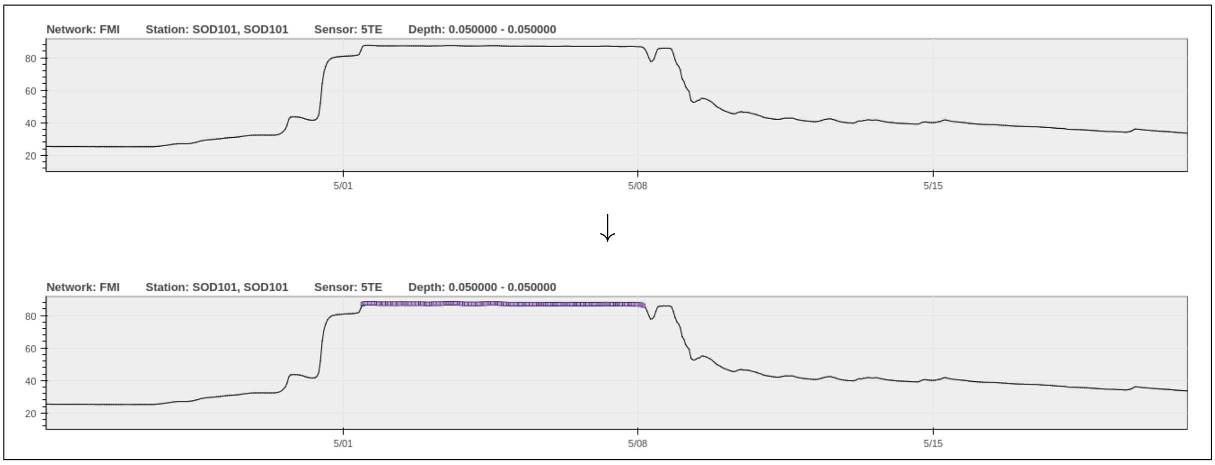


Figure 5.16: previous under flagged saturated plateau

Like in the existing methods also in the revised methods correct data values are often mistaken for saturated plateaus. Often other correct values are affected than before due to completely new conditions, but the problem is still present and according to the error matrices even more frequent than before. Since the alteration of the condition had such a positive impact on the missing flagging of erroneous data, the increased over flagging is acceptable. Figure 5.17 shows a timeseries were the improvement has a positive affect on the over flagging.

Saturated plateaus at a very high soil moisture level often have variations that cannot be accepted without over flagging correct data too much. Some plateaus are therefore only partly flagged where the boundary for the variation holds (Figure 5.18).

The saturated plateaus were most difficult to improve. Much over flagging had to be accepted to avoid plateaus that are not marked as erroneous, which would be much more severe for a validation of data sets that were retrieved by other means. The correlation of the data sets without excluding the plateaus would be wrongly

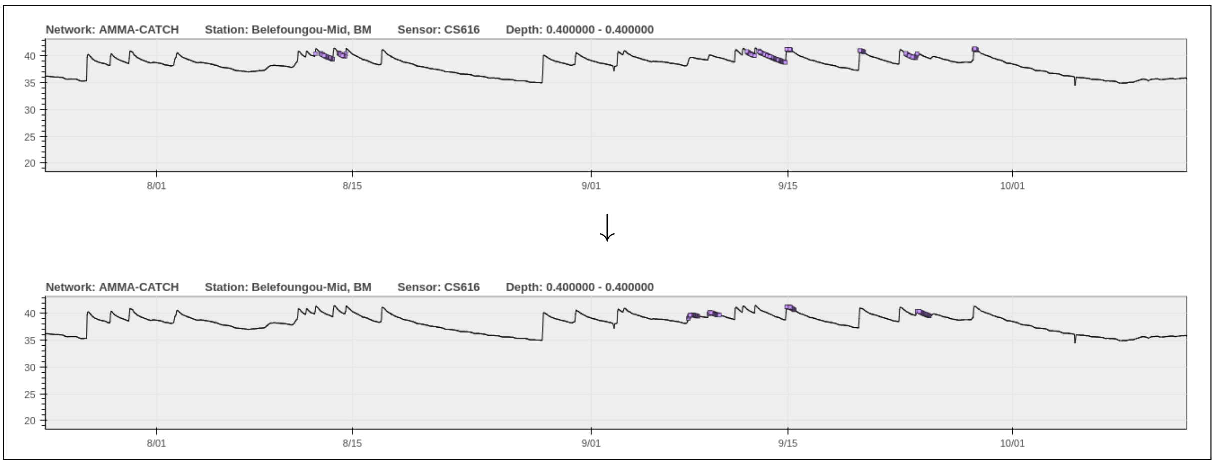


Figure 5.17: previous over flagged saturated plateaus

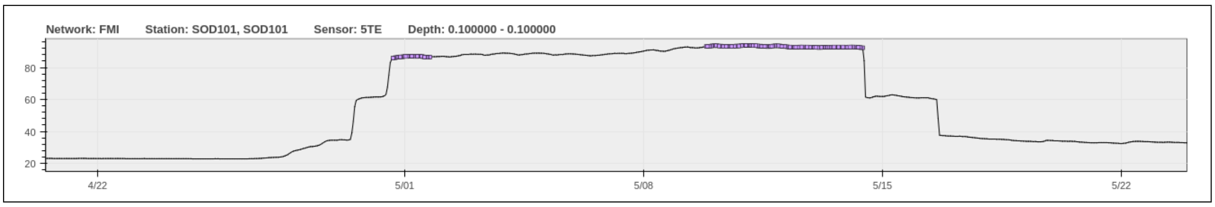


Figure 5.18: example for a saturated plateau that is still not completely flagged

decreased.

5.2.4 Suspicious Values Around Missing Values

The flag suspicious values around missing values is a valuable addition to the flagging system. Figure 5.19 shows an example of a timeseries with many of those occurrences. Those obviously bad data values look similar to spikes, but they cannot be detected as such because there are no neighboring values they can be compared to. The new flags solve this problem.

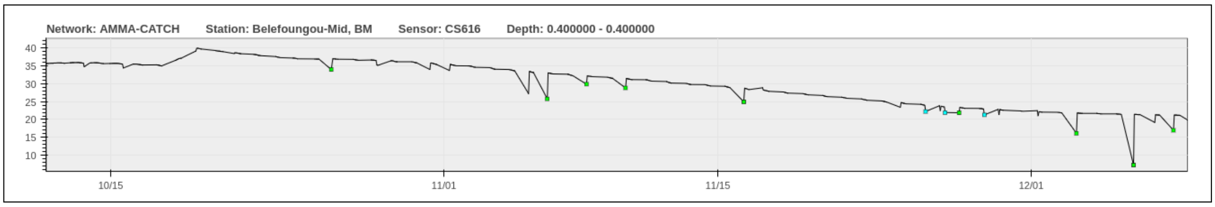


Figure 5.19: example for suspicious values Around missing values

The figure also shows a bit of an under flagging, when the difference to the existing spectrum is not high enough. The whole data set also shows a continuous drop of soil moisture, which results in a relatively high gradient, where the boundary variance sometimes does not hold. Loosening these thresholds would result in an over flagging of correct data sets.

Over flagging of the new flags could only be detected at rather noisy low soil moisture readings, where the manual flagging is rather subjective 5.20. Therefore that effect is acceptable.

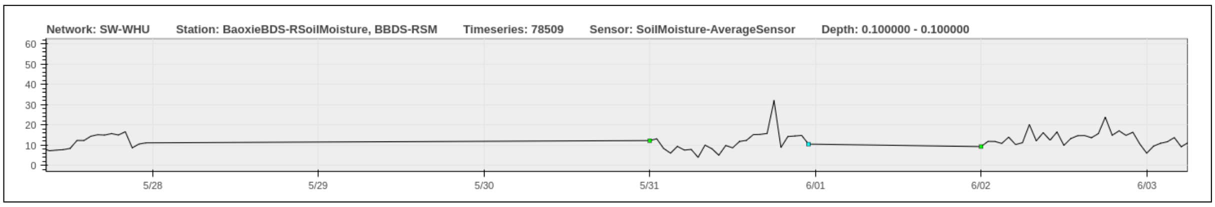


Figure 5.20: example for a correct value that is flagged as a suspicious value around missing values

5.2.5 Severe Soil Moisture Drop

This flag is a rather simple, but valuable addition. Many suspicious data series that cannot be detected with one of the introduced algorithms, which are too specific, are at least partly flagged with a sever soil moisture drop (Figure 5.21).

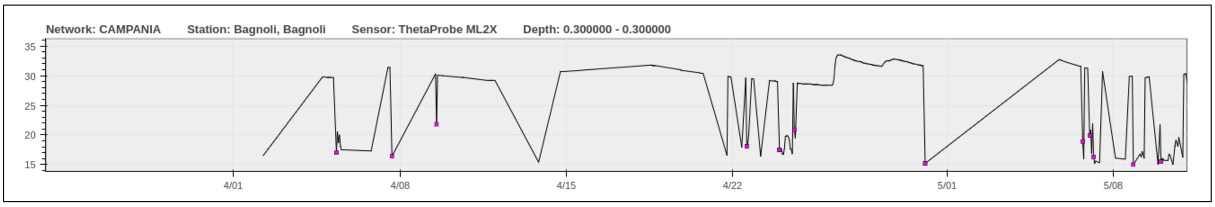


Figure 5.21: sever soil moisture drops in a suspicious data series

Soil moisture drops, which seem very high to an observer that are not flagged as such mostly occur after a sensor drop out (Figure 5.22). The linear vertical line indicates that there are no measurements in that time interval. The figure shows that those readings that exist seem realistic and are most likely connected with a soil moisture drying process. Therefore such occurrences are not flagged.

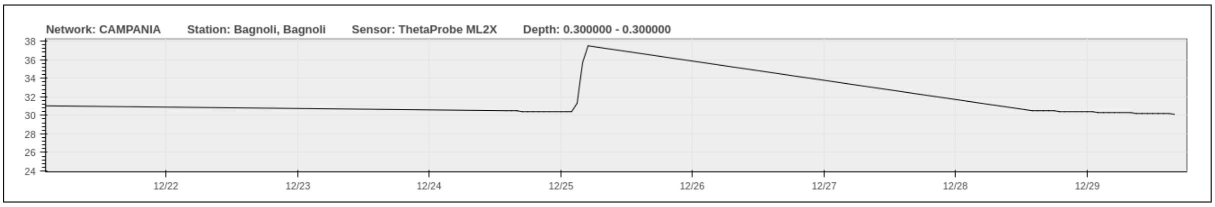


Figure 5.22: example for a severe soil moisture drop that is not flagged

The over flagging also only occurs at noisy low level soil moisture readings (Figure 5.23). As they are in any case suspicious, those results are acceptable.

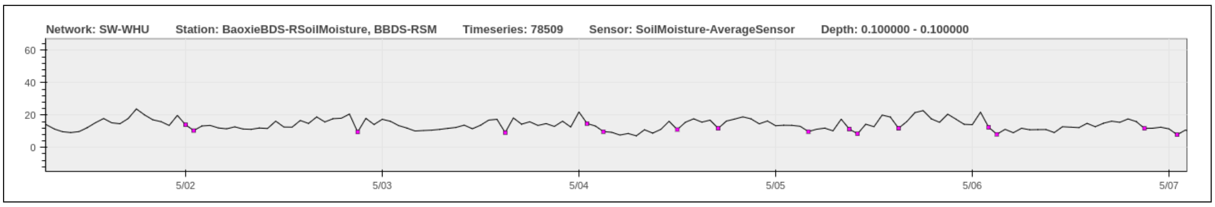


Figure 5.23: example for a correct data set that is flagged with a severe soil moisture drop

5.2.6 Alternating Values

Alternating values occur due to severe sensor or data logger malfunctions (Figure 5.24). Some parts of such a set of erroneous data values are detected by other flags, but not the whole spectrum. Such undetected values would have a very bad influence on a validation with other data sets. Therefore this flag is a very useful addition for the quality assessment.

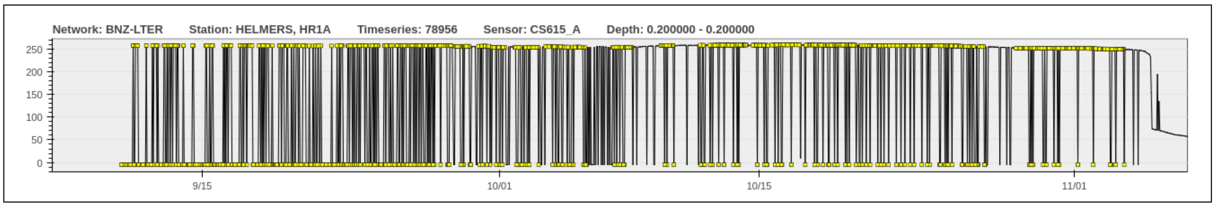


Figure 5.24: example for alternating values

Such a set of erroneous data also occurs within a constant drop or rise of soil moisture (Figure 5.25). This case is not flagged due to the high variance. Alternating values can also hardly be detected with the set of high values and the set of low values showing no significant difference. To allow one of these cases to be flagged would result in a huge unacceptable over flagging of correct data sets. Therefore this under flagging is acceptable.

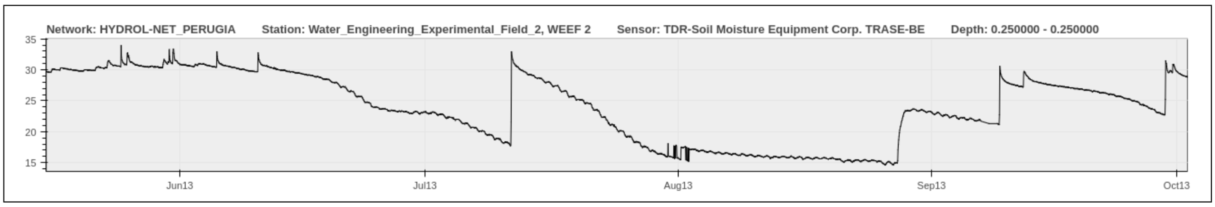


Figure 5.25: example for alternating values that are not flagged

Data sets might seem over flagged if they do not really look similar to the specific behavior. In Figure 5.26 the lower soil moisture values are possibly correct data sets, but since there is obviously something suspicious about this timeseries those over flagging is acceptable.

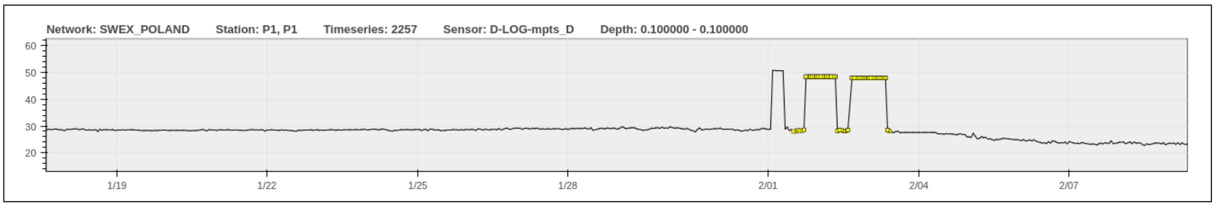


Figure 5.26: example for correct data values that are flagged as alternating values

5.2.7 Constant Values

Constant values are very unnatural in soil moisture readings. Soil moisture observations show yearly variations, influences of radiation, temperature and very high variations affected by precipitation and snow melting processes. Constant values often occur in winter when the soil is frozen, but also if the air or soil temperature is already above zero degrees, which would not be detected by the geophysical based flags. The sensors often need time to react again and provide reasonable data. Those and generally suspicious constant data values can be detected with this new flag. (Figure 5.27)

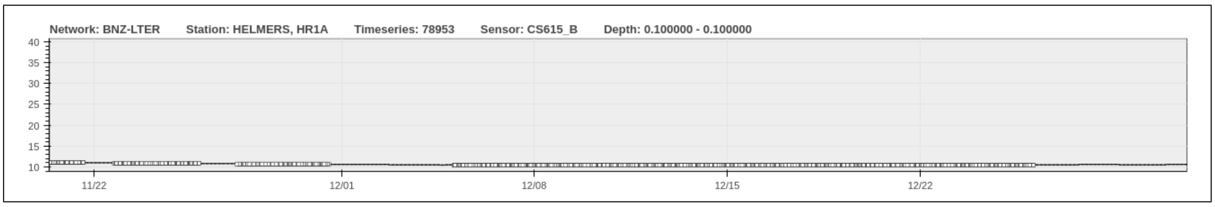


Figure 5.27: example for constant values

5.2.8 Highly Flagged Spectrum

The last flag is an appropriate completion of the ISMN spectrum based quality flags. It fills gaps of under flagged erroneous observations. In 5.24 some values of the alternating spectrum were not flagged, because they were not close enough to the mean value of the others. This problem is fixed with this last flag as it flags values if they lie in a spectrum were more than half of the datasets in 24 hours before and after are flagged. (Figure 5.28)

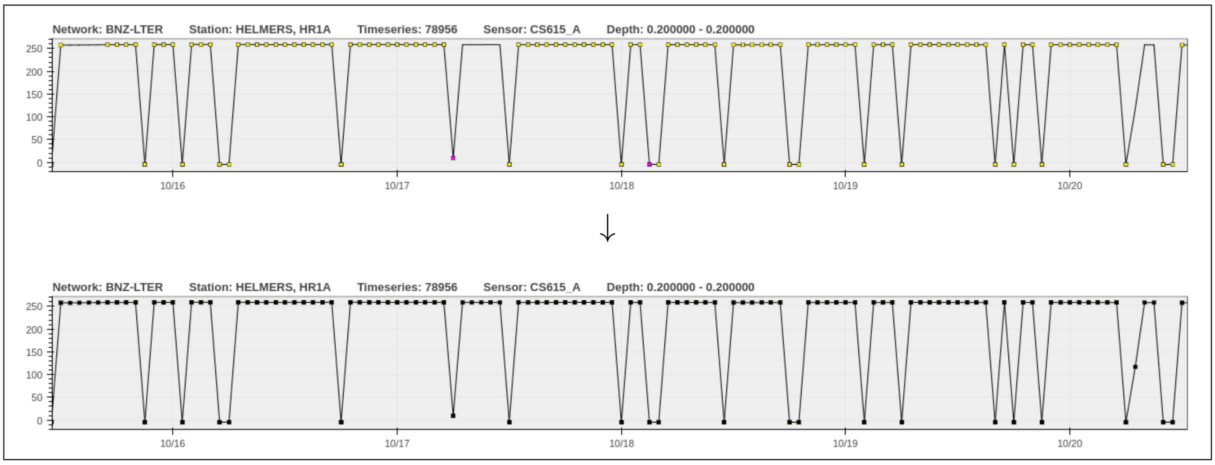


Figure 5.28: example for a filled flagging gap

5.3 Impact of the Existing and Revised Quality Flags on Validation Purposes

Soil moisture data derived from satellite data is most commonly validated by calculating the correlation with in situ measurements. The in situ observations are therefore taken as reference and for the calculations assumed to be without error, which is not the case with observed data sets. To reduce the negative impact of erroneous in situ data on the correlation coefficient the ISMN quality flags can be used to exclude the detected erroneous in situ data for a validation. Using only the observations that were found to be good by the ISMN algorithms, the correlation should increase, if the comparing data set is reasonable.

To test the effect of the ISMN quality flags, all ISMN data sets from table 5.1, which were also used for the evaluation, were validated with the corresponding GLDAS data. The GLDAS version 1 data set, provided by NASA, with a spacial resolution of 0.25 degree and temporal resolution of three hours was chosen for the validation. GLDAS data is a model data set that was selected, because it has been a reliable source for the geophysical quality flags (Table 3.3).

The correlation was calculated with the Pearson correlation coefficient for three different data set combinations. Once with all in situ measurements, once after excluding all observations that were flagged with the existing spectrum based methods and once analog for the revised spectrum based flags. The overall correlation of all 44 data sets resulted to:

- using all in situ observations: r = 0.26
- excluding flagged data sets by existing method: r = 0.27
- excluding flagged data sets by revised method: r = 0.30

The overall correlation is rather low, but the specific data sets were chosen, because they show many erroneous observation. Some of those data sets even showed a negative correlation and therefore have a very negative influence on the over all correlation. Also the grid resolution of 0.25 degree of the GLDAS data set is very low and therefore the data of a specific in situ station is not appropriately represented by with the mean grid value of the corresponding GLDAS data. More important is the difference between the correlations, which is a success.

The types of error are also noticeable in the correlation coefficient for specific data sets. The differences of the correlations are higher for a spectrum with plateaus or constant values, because more observations are excluded from the data set, e.g. the network HOBE (Figure 5.29). The graphs show the data set of the network HOBE that was chosen for the evaluation, with the flagged data set by the existing flags at the top and the revised flags at the bottom.

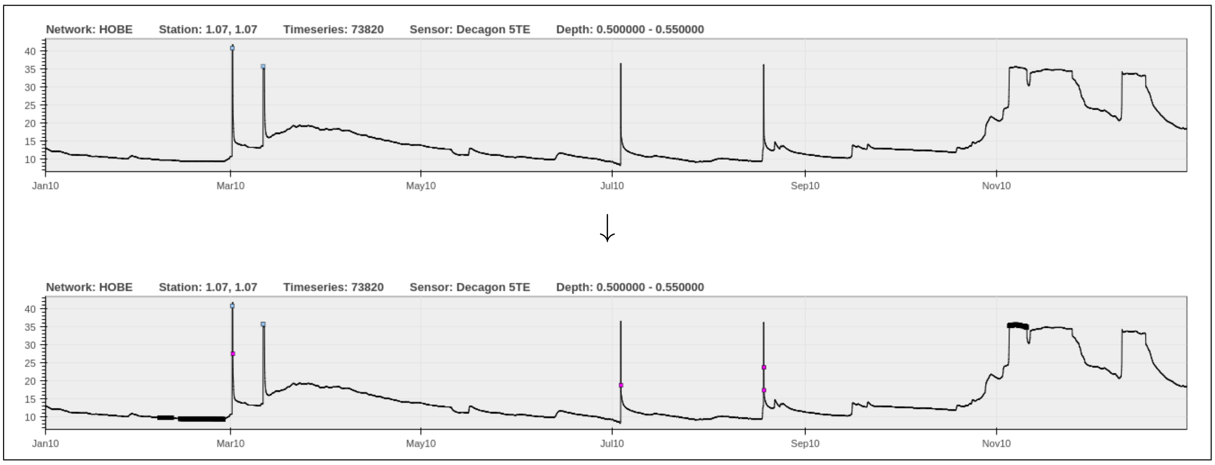


Figure 5.29: test data set of the network HOBE

For the chosen data set of the network HOBE, where a saturated plateau and constant values could be detected, the correlation coefficients are:

- using all in situ observations: r = 0.39

- excluding flagged data sets by existing method: $\mathrm{r} = 0.39$
- excluding flagged data sets by revised method: $\mathrm{r} = 0.41$

In a timeseries with errors that last only for one measurement, e.g. jumps or breaks, the difference of the correlation coefficients is much smaller.

6 Conclusion

The existing quality control methods for the ISMN are very sophisticated and especially the fundamental ideas for the algorithms are very advanced. Nevertheless, there is much room for improvement, due to the very special behavior of soil moisture data. Much more data sets are included in the ISMN data base since the development of the existing algorithms, and it was found that some error definitions need adjustments and new error types could be identified. It transpired that the best way to improve the existing algorithms was to loosen the thresholds in most of the formulas, but therefor use all characteristics the derivatives have to offer to add other conditions to avoid over flagging.

The original quality flags of the data providers have a more basic approach than those of the ISMN, but it turned out that this is also a helpful addition to flag erroneous observations that do not follow conditions of specific algorithms like breaks and spikes, e.g. a sudden soil moisture drop.

All the spectrum based quality flags could be improved and six new flags were added based on new ideas of the author and approaches of ISMN data providers. The improved flags are mostly based on empirical tests and visual inspection with a data viewer specifically developed for this thesis to examine the effect of conditions for the error detection. The statistics show that the functionality of the new flags improved evidently and the new flags are a valuable addition to avoid undetected erroneous data.

The new flags will be implemented into the ISMN processing chain as soon as possible.

An idea that was not followed for this thesis is to compare observations of sensors that are installed at the same place but at different depths, but variation of soil moisture data is highly reduced at deeper depths. Also, stations that are close to each other could be compared to examine the reasonableness of the data. For both approaches an over flagging of good data seems to be unavoidable, since the data sets often differ even if they are both correct. Additionally, it would be very complex to determine the correct and the erroneous observations.

Glossary

break A significant break in a 12 hour spectrum with elsewise continoues data sets. 14, 16, 67

network All in-situ stations of a single data provider. 6, 8

original quality flags Quality control procedure of ISMN data providers. 1

plateau Minimum 12 hours lasting continues measurements with a break before and after. 14, 19, 20, 38

savitzky-golay filter A data smoothing filter for data points which may also give a derivative of the smoothed data set. 14

spectrum Measurements of a timeseries of a specific time range. 33, 61

spike A single measurement that significantly out lies other measurements in a 12 hour spectrum. 14, 15, 33, 34

timeseries Data set with a fixed station, sensor and depth. 3, 4, 9, 14, 16, 20, 28, 38, 39, 42, 45, 49, 53, 57, 59, 63, 67

Acronyms

AACES Austrialian Airborne Cal/Val Experiment for SMOS.

ARM Atmospheric Radiation Measurement climate Research Facility.

AWDN Automated Weather Data Network.

BIEBRZA_S-1 Instytut Geodezji i Kartografii.

BNZ-LTER Bonanza Creek Long-Term Ecological Research. 23

CEOP Code of Practice. 9

COSMOS COsmic ray Soil Moisture Observing System. 3

CTP_SMTMN Central Tibetan Plateau Soil Moisture and Temperature Monitoring Network.

ESA European Space Agency. 1

FMI Sodankyla-Pallas Satellite Data Calibration and Validation Site.

GLDAS Global Land Data Assimilation System. 9, 31, 61, 62

GPS Global Positioning System. 3

HOBE Hydrological Observatory.

HSC_SELMACHEON Hydrological Survey Center.

ICN Illinois Climate Network.

IIT_KANPUR Indian Institute of Technology Kanpur.

in situ ground based. 1, 3, 4, 6, 8, 9, 12, 61, 62

iRON Roaring Fork Observation Network.

ISMN International Soil Moisture Network. xiii, 1, 3, 4, 5, 8, 9, 10, 12, 14, 18, 21, 22, 23, 25, 26, 28, 30, 31, 33, 48, 52, 60, 61, 65, 67

METEROBS Met European Research Observatory.

MOL-RAO Lindenberg Meteorological Observatory - Richard Aßmann Observatory. 24, 25, 26

NaN Not a Number. 26

NASA National Space Agency. 9, 61

NASMD North American Soil Moisture Databse. 9, 31

NLDAS North American Land Data Assimilation System. 31

P precipitation.

PBO_H2O Plate Boundary Observation. 3, 8

RSMN Romanian Soil Moisture Network.

SD Snow Depth.

SM Soil Moisture.

SMOS Soil Moisture and Ocean Salinity.

SNOTEL SNOwpack TELemetry.

SOILSCAPE SOIL moisture Sensing Controller And oPtimal Estimator. 27, 28

ST Soil Temperature.

SU Soil Suction.

SWEQ Soil Water Content.

SWEX_POLAND Soil Water and Energy eXchange - Poland.

TA Air Temperature.

TDR Time-Domain Reflectometer. 3

TERENO Terrestrial Environmental Observatories. 28

TS Soil Temperature.

TSF Surface Temperature.

UDC_SMOS Upper Danube Catchment SMOS validation side.

UMBRIA "Civil Protection Functional Centre and Research Institute for Geo-Hydrological Protection.

USCRN U.S. Climate Reference Network. 29, 30

UTC Coordinated Universal Time. 4

VAS Valencia Anchor Station.

WSMN Wales Soil Moisture Network.

Bibliography

[1] Jesse E Bell, Michael A Palecki, C Bruce Baker, William G Collins, Jay H Lawrimore, Ronald D Leeper, Mark E Hall, John Kochendorfer, Tilden P Meyers, Tim Wilson, et al. Us climate reference network soil moisture and temperature observations. *Journal of Hydrometeorology*, 14(3):977–988, 2013.

[2] WA Dorigo, Wolfgang Wagner, Roland Hohensinn, Sebastian Hahn, Christoph Paulik, Angelika Xaver, Alexander Gruber, Matthias Drusch, Susanne Mecklenburg, P van Oevelen, et al. The international soil moisture network: a data hosting facility for global in situ soil moisture measurements. *Hydrology and Earth System Sciences*, 15(5):1675–1698, 2011.

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

[4] Alexander Gruber, WA Dorigo, S Zwieback, A Xaver, and W Wagner. Characterizing coarse-scale representativeness of in situ soil moisture measurements from the international soil moisture network. *Vadose Zone Journal*, 12(2), 2013.

[5] Kristine M Larson, Eric E Small, Ethan D Gutmann, Andria L Bilich, John J Braun, and Valery U Zavorotny. Use of gps receivers as a soil moisture network for water cycle studies. *Geophysical Research Letters*, 35(24), 2008.

[6] Mahta Moghaddam, Dara Entekhabi, Yuriy Goykhman, Ke Li, Mingyan Liu, Aditya Mahajan, Ashutosh Nayyar, David Shuman, and Demosthenis Teneketzis. A wireless soil moisture smart sensor web using physics-based optimal control: Concept and initial demonstrations. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 3(4):522–535, 2010.

[7] Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

[8] Angelika Xaver. Automated quality control procedures for the international soil moisture network. Diploma thesis, TU Wien, 2015.

[9] Youlong Xia, Trent W Ford, Yihua Wu, Steven M Quiring, and Michael B Ek. Automated quality control of in situ soil moisture from the north american soil moisture database using nldas-2 products. *Journal of Applied Meteorology and Climatology*, 54(6):1267–1282, 2015.

[10] Steffen Zacharias, Heye Bogena, Luis Samaniego, Matthias Mauder, Roland Fuß, Thomas Pütz, Mark Frenzel, Mike Schwank, Cornelia Baessler, Klaus Butterbach-Bahl, et al. A network of terrestrial environmental observatories in germany. *Vadose Zone Journal*, 10(3):955–973, 2011.

[11] Fedro S Zazueta and Jiannong Xin. Soil moisture sensors. *Soil Science*, 73:391–401, 1994.

[12] Marek Zreda, Darin Desilets, TPA Ferré, and Russell L Scott. Measuring soil moisture content non-invasively at intermediate spatial scale using cosmic-ray neutrons. *Geophysical Research Letters*, 35(21), 2008.

[13] Marek Zreda, WJ Shuttleworth, Xubin Zeng, Chris Zweck, D Desilets, T Franz, and R Rosolem. Cosmos: the cosmic-ray soil moisture observing system. *Hydrology and Earth System Sciences*, 16(11):4079–4099, 2012.