Automated Quality Control Procedures for the International Soil Moisture Network

ANGELIKA XAVER[1], WOUTER DORIGO[1] & WOLFGANG WAGNER[1]

Zusammenfassung: In situ Bodenfeuchtebeobachtungen sind essentiell, um modellierte und fernerkundungsbasierte Bodenfeuchteprodukte zu evaluieren und zu kalibrieren. Daher sind aussagekräftige Qualitätsmaße für in situ Bodenfeuchtemessungen von höchster Relevanz. Basierend auf den Daten des International Soil Moisture Network (ISMN), werden komplexe automatisierte Verfahren der Qualitätskontrolle präsentiert, um Ausreißer, Sprünge und Plateaus mithilfe der Form von Bodenfeuchte-Zeitreihen zu detektieren. Mehrere Bedingungen werden definiert, um diese fehlerhaften Ereignisse durch Untersuchung der ersten und zweiten Ableitungen, berechnet mit dem bekannten Savitzky-Golay Filter, zu identifizieren. Die Performance der vorgestellten Qualitätskontrolle wird durch einen Vergleich der automatisiert detektierten Resultate mit manuell markierten Daten, basierend auf einer Auswahl von 40 Bodenfeuchtedatensätzen des ISMN, evaluiert.

1 Introduction

Satellite-derived soil moisture products represent the most promising source of global and long-term soil moisture data availability. However, in situ soil moisture measurements are still crucial for evaluating and calibrating satellite-derived and model-based soil moisture products. Despite being measured directly in the soil, in situ measurement time series typically contain a large number of artefacts such as outliers and breakpoints. Thus, knowledge about the quality of in situ soil moisture measurements is fundamental when using them to assess the reliability of any satellite or model soil moisture product (e.g. DORIGO et al. 2015; ALBERGEL et al. 2012).
This paper presents novel automated quality control procedures for in situ soil moisture time series. The plausibility of soil moisture measurements is evaluated by examining the shape of soil moisture time series by means of its first and second derivative, calculated by the widely-known Savitzky-Golay filter. The advanced automated time series-based methods have been developed and tested using soil moisture data from the International Soil Moisture Network (ISMN) (DORIGO et al. 2011a,b), the largest data bank of in situ soil moisture data worldwide.

2 Data

2.1 International Soil Moisture Network

The International Soil Moisture Network (ISMN; http://ismn.geo.tuwien.ac.at/; DORIGO et al. 2011a,b), has been developed and operated by the Department of Geodesy and Geoinformation (GEO) of the TU Wien since 2010. The ISMN acts as a data repository for ground-based soil

[1] Technische Universität Wien, Department Geodäsie und Geoinformation, Forschungsgruppen Photogrammetrie und Fernerkundung, Gußhausstraße 27-29, A-1040 Wien,
E-Mail: [Angelika.Xaver, Wouter.Dorigo, Wolfgang.Wagner]@geo.tuwien.ac.at

moisture measurements, with the primary goal of providing reference data for calibrating and validating remotely-sensed soil moisture missions (e.g. SMOS, SMAP). The ISMN collects soil moisture observations from various networks distributed all over the globe. After running through a fully automated processing chain, where the observations are harmonized in terms of unit, temporal resolution and data format, the data is stored in a database and becomes available to the public through a web portal.

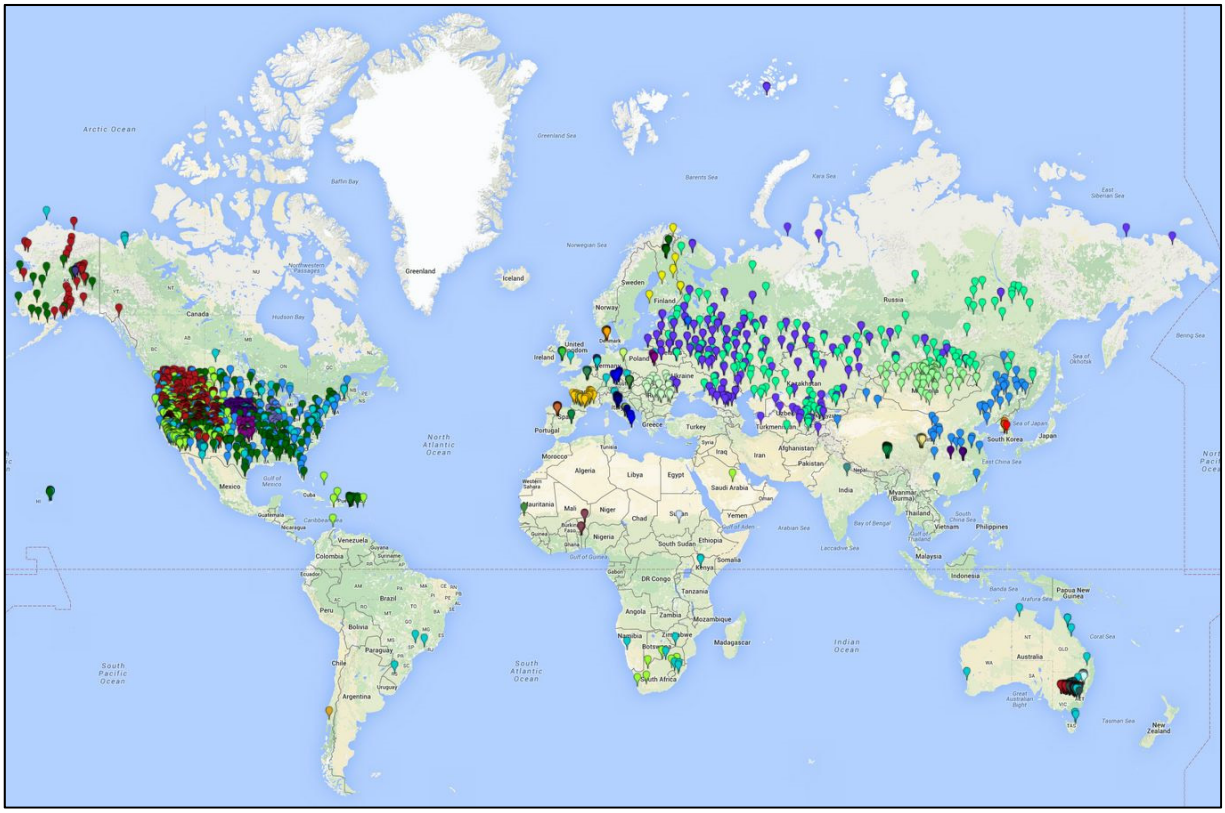

Fig. 1: Stations available at the online platform of the ISMN (status April 2016)

Currently, the ISMN stores the data of 53 networks consisting of 2144 stations (Figure 1). While most datasets are updated on an irregular basis, a few networks provide their data in near-real-time (NRT). These datasets are automatically downloaded, processed and written into the database. Due to the large amount of data processed with each update it is infeasible to perform an operator-based manual quality control of the data. Thus, automated quality control procedures within the ISMN are important in order to provide reliable soil moisture measurements to its users. Furthermore, not every network applies quality control procedures to their data and if quality control mechanisms exist, they are not consistent. Thus, the importance of harmonized quality control procedures, which can be applied to all networks consistently, is evident.

2.2 Soil moisture characteristics

The difficulty in defining quality principles for in situ soil moisture measurements lies in the characteristic shape of a soil moisture time series, which is represented by alternating phases of

wetting and drying (HILLEL 1998). Precipitation events may result in a severe rise of soil moisture within only one or a few hours, while the drying process of soil proceeds slowly, resulting in an inverse exponential shape (Figure 2). These specific properties inhibit the use of typical outlier detection algorithms, which would flag most of the natural rises of soil moisture caused by rain events as erroneous.

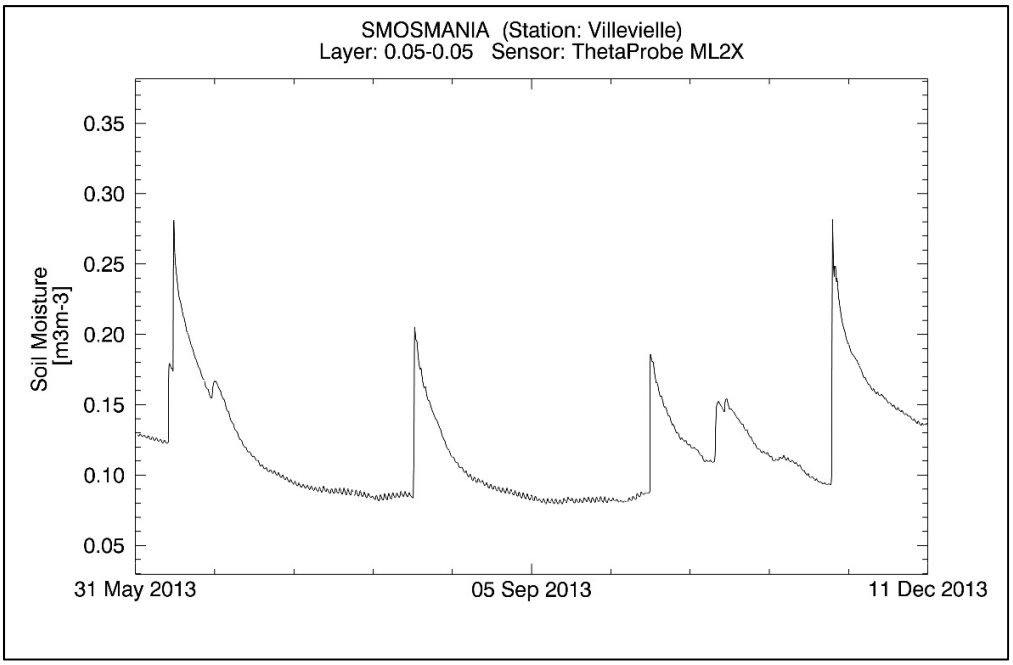

Fig. 2: Example of characteristic wetting and drying phases of an in situ soil moisture time series

Furthermore, the identification of outliers is complicated by the complexity of the signal, e.g. consecutive precipitation events may result in overlapping rises of soil moisture, variations induced by daily temperature cycles may exist (DORIGO et al. 2011b; ROBINSON et al. 2008), or random noise can be present within the soil moisture observations.
In fact, soil moisture regimes may vary strongly depending on the prevalent climate, vegetation and soil composition. Ideally, quality control algorithms should be able to cope with these unavoidable natural and artificial phenomena. In reality, a trade-off has to be made to identify suspicious measurements without over-flagging natural events.

2.3 Potential erroneous events within soil moisture readings

A variety of suspicious measurements can occur caused by malfunction or irresponsiveness of the sensor, or a connection problem while writing on the data logger. The resulting erroneous shapes within the soil moisture time series can be generalized into three categories:

- Spikes: A “spike” is as an outlier of a single measurement which lies significantly above or beneath the preceding and succeeding measured soil moisture values (Figure 3, top left). Thus, a “spike” lasts only for a single unit of time and can appear in both positive and negative directions.

- Breaks: A “break” is a sudden, from one timestamp to the next, rise (“positive break”) or fall (“negative break”) in the measurement time series (Figure 3, bottom left). A sudden drop of the level of soil moisture cannot occur in reality, whereas an instant rise in the amount of soil moisture can either have a natural cause, i.e. as the result of a precipitation event, or be an artefact. Problematic is the fact that breaks may lead to a general offset (bias) of the soil moisture signal.
- Constant Values: Two different kinds of constant values may occur. First, constant values can occur if consecutive precipitation events result in high soil moisture readings, where the sensor is not able to represent any further variation of water in the soil beyond that level. This is referred to as “saturated plateau” (Figure 3, top right). Second, constant values may occur after a sudden drop (“negative break”) of soil moisture, for instance caused by an energy supply problem or when the soil is frozen. This event will be referred to as “low level plateau” (Figure 3, bottom right).

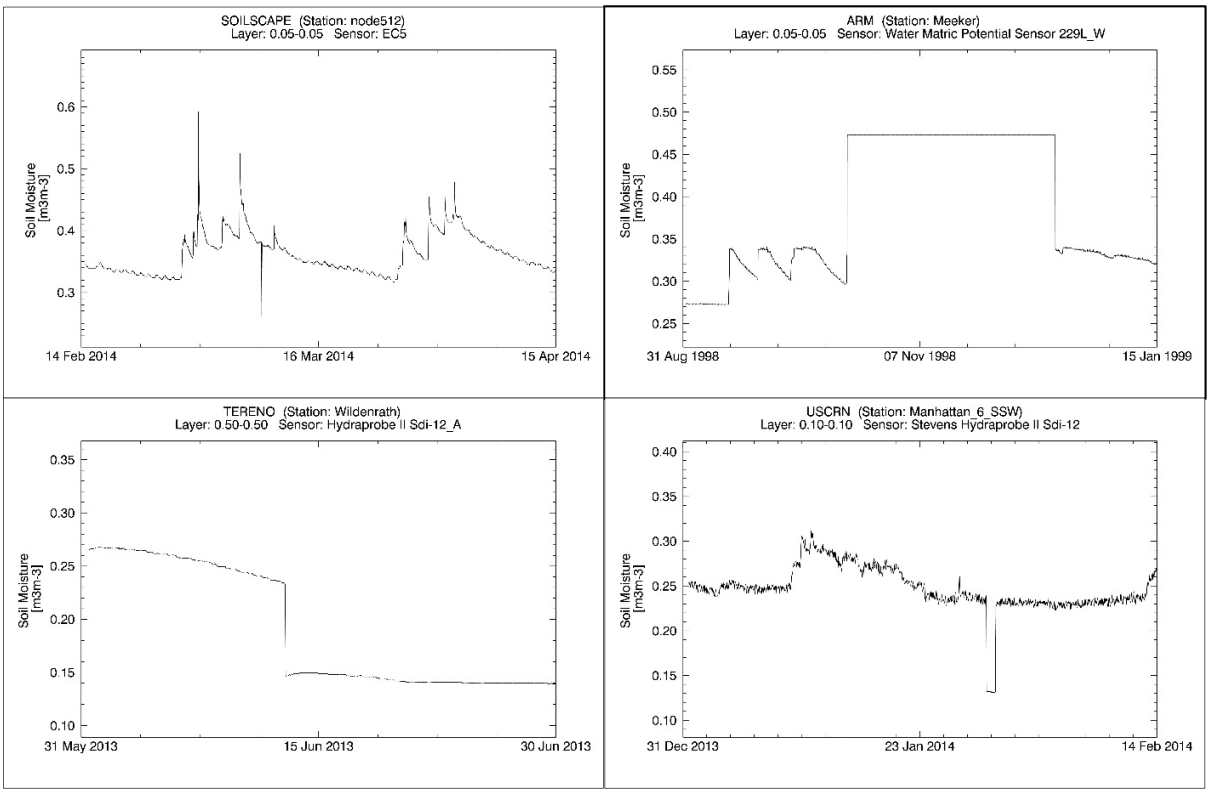

Fig. 3: Examples of erroneous events within soil moisture time series. From top left to bottom right: (negative) spike, saturated plateaus, negative break and low level plateaus.

3 Methods

3.1 Savitzky-Golay Filter

The detection of spurious events is based on the first and second derivative of soil moisture time series, which are calculated by using a Savitzky-Golay filter (SAVITZKY & GOLAY 1964). The advantage of this method is not only the ability to get a smoothing filter and derivatives with the

same equation, but also to preserve higher moments (FLANNERY et al. 1992). Short periods of missing values can be handled well by the Savitzky-Golay filter (EILERS 2003). In addition, the computation is simple, as the 'smoothing coefficients' are applied to the data by a convolution, and therefore fast, i.e., suited for NRT processing.
The Savitzky-Golay filter was parametrized with a small symmetrical filter width of three data points and a second degree polynomial.

3.2 Detection Algorithms

Based on the characteristic appearance of each type of suspicious events within the first and second derivative the following conditions were developed through empirical investigation. More details, i.e. all equations, can be found in DORIGO et al. (2013) and XAVER (2015).

3.2.1 Spikes

A spike is detected if the following conditions are fulfilled:

1. The soil moisture signal increases or decreases by at least 15%, which corresponds to approximately three times a typical maximum sensor uncertainty.
2. A spike within the soil moisture time series results in a spike in the second derivative surrounded by smaller peaks in opposite sign (Figure 4, bottom). These surrounding peaks have to be of similar size, thus their ratio is approximately one.

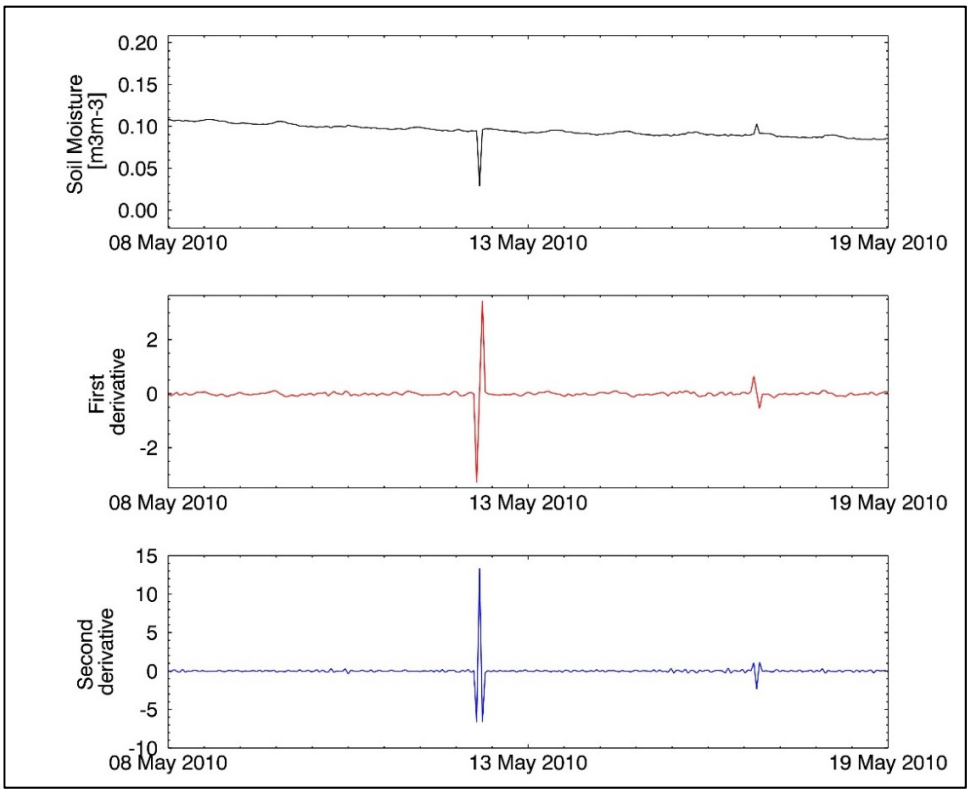

Fig. 4: Shape of first (middle) and second derivative (bottom) for a spike within the soil moisture time series (top) for Station Eulo of network OZNET measured with a Stevens Hydra Probe in a depth from 0.00 to 0.05m

3. The coefficient of variation is applied to the soil moisture time series to an interval of 12 hours before and 12 hours after the potential spike, but excluding the measurement of the spike itself. Its value has to be below one. With this condition an over-flagging in the case of noisy data is avoided.

3.2.2 Breaks

A break is detected if the following conditions are fulfilled:

1. The relative change of soil moisture with respect to the previous time step has to be at least 10%, whereas the absolute amount of change has to be at least 0.01m^3m^{-3} to avoid over-flagging.
2. A characteristic strong negative (positive) change within the first derivative can be observed when a negative (positive) break appears in the soil moisture time series. This first derivative jump has to be at least ten times smaller (larger) than the arithmetic mean of a 25-hours period of the first derivative centered at the potential break.
3. A negative (positive) break leads to a pronounced negative (positive) second derivative value at the time step before the break, followed by a large positive (negative) value a single time step after the break. These two values have approximately the same magnitude, thus their ratio must be close to unity.

3.2.3 Plateaus

Soil moisture observations are flagged as saturated plateaus if the following conditions are fulfilled:

1. The variation of soil moisture readings over a period of at least 12 hours does not exceed 1% of a typical average sensor uncertainty of 0.05 m^3m^{-3}.
2. The characteristic strong rise of soil moisture at the beginning and the sudden drop at the end of the saturated plateau are represented in the first derivative through a sharp positive peak at the beginning, and a strong negative peak at the end of the plateau. Those peaks have to exceed certain threshold values (DORIGO et al. 2013, XAVER 2015).
3. Saturated plateaus consist of the highest recorded soil moisture values of the time series, thus, the level of soil moisture during the plateau has to be at least 95% of the maximum value observed during the whole measurement period.

Soil moisture observations are flagged as low level plateau if the following conditions are fulfilled:

1. A negative break, as described above, is detected.
2. As long as the coefficient of variation stays below 0.01 the measurements following the negative break are identified as low level plateau.

4 Results and Discussion

The performance of the automated quality control procedures was evaluated by comparing the flags obtained from the proposed quality control procedures to a visual classification of occurring erroneous measurements. For this purpose, a selection of 40 datasets from 19 different networks

was drawn from the ISMN and inspected visually. The result of the flagging performance is presented in Table 1.
The overall percentage of flagged data is less than 10% of all readings. As some of the erroneous events represent only one time step (e.g. spikes and breaks) this low outcome does not surprise. Depending on the event, the detection accuracy varies between 42 and 92%. The percentage of correct observations detected as erroneous measurements, i.e. the number of false alarms, lies under 2%.

Tab. 1: Results of the flagging performance (all values given in percentage)

	Erroneous measurements		Correct measurements		
Flagging results	‚Erroneous'	‚Correct'	‚Erroneous'	‚Correct'	Flagged observations
Spikes	80.0	20.0	0.0	100.0	0.0
Positive breaks	41.5	58.5	0.0	100.0	0.0
Negative breaks	57.4	42.5	0.0	100.0	0.0
Low level plateaus	59.6	40.4	1.4	98.6	2.8
Saturated plateaus	92.2	7.7	1.6	98.4	6.3

In general, the quality control algorithms are working very well (see Figure 5 for examples). Reasons for the failure of the flagging procedures can be summarized as follows:

- Existence of missing values: Single missing values can be overcome by the introduced algorithms, but multiple missing values or whole periods of missing data lead to a disturbance of the Savitzky-Golay filter and the characteristic shapes in the first and second derivatives.
- Imperfect definition: Some thresholds had to be defined for the different quality control procedures, e.g. for identifying saturated plateaus, to detect as many erroneous events as possible while avoiding over-flagging of correct data. Of course, real data does not always comply with these thresholds and thus some erroneous measurements cannot be detected.
- Noisy data: Noisy data causes also noisy first and second derivatives, resulting from the applied parameterization of the Savitzky-Golay filter. Therefore, the characteristics of the two derivatives which are used to identify the different suspicious measurements are superimposed by the random noise.
- Consecutive erroneous events: The effect of consecutive erroneous events is similar to that of noisy data. The characteristic shape of the derivatives is disturbed and the defined methods fail to detect erroneous measurements.

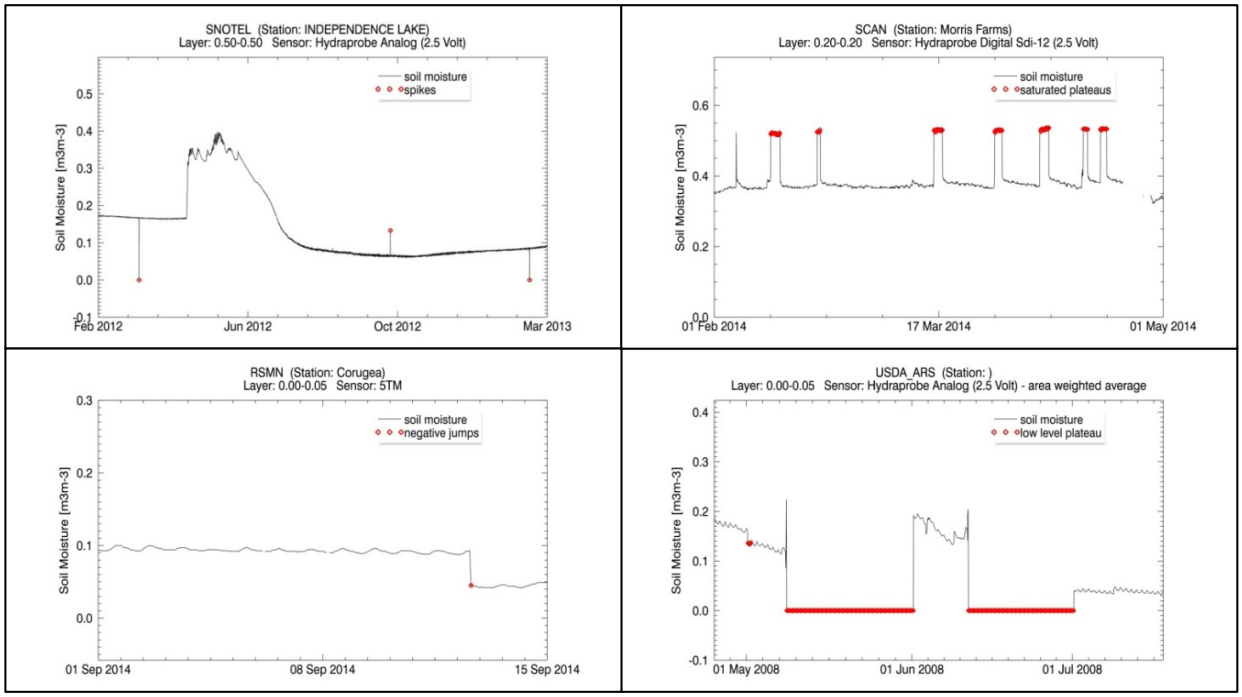

Fig. 5: Examples of flagging results obtained by the automated quality control procedures. From top left to bottom right: spikes, saturated plateaus, negative break and low level plateaus.

5 Conclusions

Automated quality control procedures cannot be expected to work perfectly. Nevertheless, an overall good percentage of erroneous measurements could be identified by the introduced methods. Moreover, the resulting number of false alarms was extremely low. Even if not all erroneous observations can be detected, the spectrum-based quality control algorithms act as a clear indicator for the quality of a soil moisture time series.

The herein described quality control procedures are implemented within the ISMN processing chain since spring 2014. For each case of the described quality detection methods (i.e., spikes, breaks, and plateaus) separate flags are defined. The flags are attached to the soil moisture measurements and provided to all users of the ISMN in addition to the actual measurements, which themselves remain unchanged. Thus, the quality control procedures will help to improve the reliability of validation studies based on in situ soil moisture observations on the one hand, and will help to identify problems which may occur at the measurement sites on the other hand.

6 References

This paper is based on the diploma thesis 'Automated Quality Control Procedures for the International Soil Moisture Network' (XAVER 2015).

ALBERGEL, C., DE ROSNAY, P., GRUHIER, C., MUNOZ-SABATER, J., HASENAUER, S., ISAKSEN, L., KERR, Y. & WAGNER, W., 2012: Evaluation of remotely sensed and modelled soil moisture products using global ground-based in situ observations. Remote Sensing of Environment **118**, 215-226.

DORIGO, W., GRUBER, A., DE JEU, R., WAGNER, W., STACKE, T., LOEW, A., ALBERGEL, C., BROCCA, L., CHUNG, D., PARINUSSA, R. & KIDD, R., 2015: Evaluation of the ESA CCI soil moisture product using ground-based observations. Remote Sensing of Environment **162**, 380-395.

DORIGO, W., OEVELEN, P. V., WAGNER, W., DRUSCH, M., MECKLENBURG, S., ROBOCK, A. & JACKSON, T., 2011a: A new international network for in situ soil moisture data. Eos Transactions AGU **92** (17), 141-142.

DORIGO, W., WAGNER, W., HOHENSINN, R., HAHN, S., PAULIK, C., XAVER, A., GRUBER, A., DRUSCH, M., MECKLENBURG, S., VAN OEVELEN, P., ROBOCK, A. & JACKSON, R., 2011b: 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.

DORIGO, W., XAVER, A., VREUGDENHIL, M., GRUBER, A., HEGYIOVÁ, A., SANCHIS-DUFAU, A.D., ZAMOJSKI, D., CORDES, C., WAGNER, W. & DRUSCH, M., 2013: Global Automated Quality Control of In situ Soil Moisture data from the International Soil Moisture Network. Vadose Zone Journal **12**, 3.

EILERS, P., 2003: A Perfect Smoother. Analytical Chemistry (75), 3631-3636.

HILLEL, D., 1998: Environmental soil physics. Academic Press.

FLANNERY, B., TEUKOLSKY, S. & VETTERLING, W., 1992: Savitzky-Golay Smoothing Filters. Numerical Recipies in C: The art of scientific computating. Cambridge University Press.

ROBINSON, D., CAMPBELL, C., HOPMANS, J., HORNBUCKLE, B., JONES, S., KNIGHT, R., OGDEN, F., SELKER, J. & WENDROTH, O., 2008: Soil moisture measurement for ecological and hydrological watershed-scale observatories: A review. Vadose Zone Journal **7**, 358-389.

SAVITZKY, A. & GOLAY, M.J.E., 1964: Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry **36**, 1627-1639.

XAVER, A. 2015: Automated Quality Control Procedures for the International Soil Moisture Network. Diploma thesis, TU Wien, Austria.