Postprint: Anomaly Detection Method for Automatic Soil Moisture Observation Data

Authors: Li Cuina, Liu Tianqi

Date: 2021-08-02T00:00:00+00:00

Abstract

Targeting nationwide automatic soil moisture real-time hourly observation data, and combining instrument calibration methods, soil hydrophysical constants, and other factors, this study investigates the inherent variation characteristics of soil moisture, analyzes the error sources of anomalous data and threshold determination, and conducts quality control experiments and tests on soil moisture observation data using five methods: frequency detection, limit value detection, no-precipitation mutation detection, anomalous peak detection, and constant value detection. The results show that: (1) Automatic soil moisture anomalous data are mainly categorized into gross values, mutations, anomalous peaks, and frozen values, primarily caused by instrument failure, inaccurate determination of soil hydrophysical constants, unreasonable sensor calibration, and other such reasons. (2) Frequency detection can effectively detect erroneous data caused by instrument failure. Currently, this method has been applied to the integrated meteorological observation data quality control operational system for conducting quality control and quality assessment of nationwide real-time automatic soil moisture hourly data.

Full Text

Preamble

Anomaly Data Detection Method for In Situ Automatic Soil Moisture Observation

LI Cuina[1], LIU Tianqi[2], WU Dongli[1]

[1]Meteorological Observation Centre, China Meteorological Administration, Beijing 100081, China
[2]Inner Mongolia Meteorological Information Centre, China Meteorological Administration, Hohhot 010051, Inner Mongolia, China

Abstract: Soil moisture serves as a critical parameter in climate models, hydrological models, ecological models, and land surface process models, and represents an important indicator for studying soil erosion, crop drought monitoring, and yield prediction. Frequency domain reflectometry (FDR) technology has become one of the most promising ground-based soil moisture monitoring techniques due to its high precision, high spatiotemporal resolution, real-time capability, and all-weather operation. The meteorological department in China has established an automatic soil moisture observation network based on this method. However, multiple factors affect the accuracy of FDR sensor measurements during observation, including temperature, soil texture, calibration methods, and equipment stability and maintenance, which in turn impact the operational application of soil moisture data. To fully leverage the network data for drought monitoring and early warning, satellite product validation, and model simulation evaluation, research on quality impact factors and processing algorithms for soil moisture data is essential.

Unlike conventional meteorological observation data, soil moisture observation data exhibit extremely uneven spatial distribution and are significantly influenced by soil texture and structure, calibration equations, and other factors. Currently, no standardized operational real-time data quality control system exists. In recent years, scholars have dedicated efforts to developing automatic quality control techniques for soil moisture data. Zhang Zhifu [1] developed a quality control scheme applicable to hourly data from national automatic soil moisture stations through statistical analysis of extreme values in long-term manually observed soil moisture data. Wang Xiaodong et al. [2] explored the primary causes of errors in automatic soil moisture data through analysis of absolute errors and correlations between manual and automatic observations, preliminarily establishing an operational quality control method for Anhui Province. Wang Liangyu et al. [3] proposed a quality control early warning method for soil moisture data by analyzing relationships among automatic soil moisture data and considering soil water redistribution patterns. Wu Dongli et al. [4] examined soil moisture data using range checks, temporal variation checks, and persistence checks. Wang Jiaqiang et al. [5] introduced relationships between ground temperature and precipitation, proposing a set of quality control methods for national automatic soil moisture observation data from perspectives of threshold checks, internal consistency checks, and temporal consistency checks. While these methods predominantly employ threshold-setting approaches that are relatively simple yet effective at identifying anomalies within thresholds, Dorigo et al. [6] adopted spectral analysis methods to systematically establish new approaches for screening and eliminating abnormal values from European International Soil Moisture Network data, which subsequent scholars have extensively applied. These research achievements provide the foundation for the preprocessing and quality control scheme design for soil moisture data in this study.

This research focuses on real-time hourly observation data from national automatic soil moisture stations in China. By integrating instrument calibration

methods and soil hydrophysical constants, we investigate the inherent variation characteristics of soil moisture, analyze error sources and threshold determination for abnormal data, and conduct quality control experiments and verification using five methods: frequency detection, threshold detection, no-precipitation sudden change detection, abnormal peak detection, and constant value detection. Results demonstrate that: (1) Automatic soil moisture abnormal data primarily consist of gross errors, mutations, abnormal peaks, and frozen values, mainly caused by instrument malfunction, inaccurate determination of soil hydrophysical constants, and unreasonable sensor calibration. (2) Frequency detection can effectively identify erroneous data caused by instrument failure. Currently, this method has been applied in the Integrated Meteorological Observation Data Quality Control Business System for quality control and assessment of national real-time automatic soil moisture hourly data.

Keywords: automatic soil moisture; quality control; abnormal data; instrument principle

2.1 Variation Characteristics of Soil Volumetric Water Content

Precipitation constitutes a crucial factor affecting soil volumetric water content. Following precipitation events, soil volumetric water content increases (using 30 cm as an example). As precipitation infiltrates and replenishes soil moisture, the content rapidly decreases after reaching its peak, with soil transitioning from wet to dry and eventually stabilizing (Figure 1). Since the dielectric constant of ice is significantly lower than that of water, soil volumetric water content during freezing periods is noticeably lower than normal values. Soil volumetric water content is also notably influenced by temperature [7]; during freeze-thaw processes, it decreases with falling temperature and increases with rising temperature (Figure 2). The soil surface layer (0–10 cm) represents the interface between the entire soil profile and the external natural environment, directly affected by external factors such as precipitation and evaporation. Consequently, the coefficient of variation for soil volumetric water content is greater in surface layers compared to deeper soils [8]. Beyond the surface layer, soil volumetric water content across various depths exhibits high correlation, with adjacent layers showing extremely significant correlations and obvious interactions [9].

2.2 Analysis of Abnormal Data Error Sources

Automatic soil moisture sensors employ the frequency domain reflectometry method, utilizing high-frequency electronic oscillation circuits to measure the dielectric constant of media based on electromagnetic wave propagation velocity, thereby obtaining soil moisture. Observation data are categorized into measured values and derived values. Measured values represent the frequency or voltage directly measured by the sensor. Volumetric water content constitutes primary data calculated from measured values using parameter equations from laboratory and field calibrations. Soil weight water content, relative humidity, and effective water storage capacity represent secondary data converted from primary data through soil hydrophysical constants determined via manual measurement.

Automatic soil moisture observation depths comprise eight layers: 0–10 cm, 10–20 cm, 20–30 cm, 30–40 cm, 40–50 cm, 50–60 cm, 60–70 cm, and 70–80 cm. In practice, observation depths can be adjusted according to requirements, combined with soil texture, groundwater level, and parent material layer depth [10].

Due to sensor failure, sudden changes in soil moisture between adjacent time intervals may occur without precipitation or when precipitation is below a certain threshold (Figure 2a). During soil cracking, observed soil moisture values continuously decrease over time until approaching or reaching 0, a phenomenon that frequently occurs in clay soils [11] (Figure 2b). During continuous precipitation events, soil volumetric water content may remain unchanged or change minimally for extended periods (Figure 2c). Unreasonable sensor calibration or non-representative observation site values can cause systematic errors between measured and actual soil moisture values. Appropriate sensor recalibration can restore normal soil moisture values (Figure 2d, dashed line).

2.3 Abnormal Data Detection Methods

2.3.1 Frequency-Based Instrument Failure Detection Method (FD)

According to the observation principle of soil moisture sensors, frequencies measured in soil should fall between those measured in air and water. By investigating and analyzing these variation characteristics to determine threshold values, instrument failure-induced data errors can be detected. This study establishes the following criteria:

$$f_t^D \leq \min(f_{t,\text{air}}^D, f_{t,\text{water}}^D) \quad \text{or} \quad f_t^D \geq \max(f_{t,\text{air}}^D, f_{t,\text{water}}^D)$$

where f_t^D represents the frequency value measured by the sensor at depth D at time t; $f_{t,\text{air}}^D$ and $f_{t,\text{water}}^D$ represent frequency values measured by the sensor in water and air, respectively; and f_t^D is the measured value of the soil sensor at depth D at time t.

2.3.2 Threshold Detection for Volumetric Water Content and Relative Humidity (VD & RHD)

Based on the measurement range of soil moisture sensors and considering water retention characteristics of soil texture, thresholds are established using the volumetric water content achievable when clay (with the best water retention) is oversaturated (Equation 2). However, oversaturation caused by irrigation or heavy precipitation, along with potentially underestimated field capacity parameters, may result in short-term relative humidity values exceeding 100%. Therefore, thresholds are set using three times the standard deviation [12], with values beyond this range considered gross errors indicating erroneous data.

$$X_t^D < 60 \quad \text{or} \quad Y_t^D < 180$$

where X_t^D represents the volumetric water content at depth D at time t; Y_t^D represents the relative humidity at depth D at time t.

For example, at Saihan District in Inner Mongolia during August 2018, the second-layer soil volumetric water content showed anomalies (marked in red in Figure 3a). Verification with the station confirmed that the issue stemmed from inappropriate instrument calibration. After recalibration, the volumetric water content observations returned to normal (dashed line in Figure 3a). At Qinghai Biantan during November 2018, the first-layer soil relative humidity was abnormal (Figure 3b), fluctuating between 150%–350%. Both parameters showed obvious positive responses to precipitation, leading to a preliminary diagnosis of inaccurate soil hydrophysical constant determination. After parameter adjustment, relative humidity recovered from abnormal levels to the normal range of 39.8%.

2.3.3 No-Precipitation Sudden Change Detection (SBD)

Precipitation serves as an important variable affecting soil moisture variation. No-precipitation sudden change detection identifies abnormal values based on the consistency relationship between precipitation and soil moisture. At time t, if the cumulative precipitation is less than or equal to a critical value and satisfies Equation (3), the value is considered abnormal.

$$X_t^D > X_{t-1}^D + 2\sigma_X^D \quad \text{and} \quad P_{\text{cum}} \leq P_{\text{crit}}$$

where X_t^D is the soil volumetric water content at depth D at time t; X_{t-1}^D is the soil volumetric water content at depth D at the previous time; σ_X^D is the standard deviation of soil volumetric water content at depth D over the past 48 hours; P_{cum} is the cumulative precipitation; and P_{crit} is the critical cumulative precipitation threshold.

The critical cumulative precipitation threshold is related to observation depth and sensor precision, calculated using Equation (4):

$$P_{\text{crit}} = A \times D \times p$$

where D is the sensor observation depth (m); A is the sensor precision, typically valued at 0.05 m^3/m^3; and p is soil porosity. A and p typically use empirical values. Precipitation-based detection of abnormal soil volumetric water content values is more suitable for surface layers (0–10 cm).

2.3.4 Abnormal Peak Detection (PD)

This method addresses the issue of erroneously marking increases in soil water content caused by precipitation as errors. An algorithm based on the time series variation of soil volumetric water content and its second derivative identifies abnormal peaks. If the observation value at time t at depth D increases or decreases by at least 15% compared to the previous moment, it is considered a potential abnormal peak (Equation 5).

$$X_t^D > 1.15 \times X_{t-1}^D \quad \text{or} \quad X_t^D < 0.85 \times X_{t-1}^D$$

A second derivative is added for further detection of these potential abnormal peaks. If the ratio of the second derivative of soil volumetric water content at time t to that at time $t-1$ falls between 0.8–1.2, the value is considered normal; otherwise, it is abnormal (Equation 6).

$$0.8 < \frac{X_t''^D}{X_{t-1}''^D} < 1.2$$

The second derivative is obtained using Savitzky-Golay filtering with a 3-hour window and quadratic polynomial fitting:

$$X_t''^D = X_{t-1}^D - 2X_t^D + X_{t+1}^D$$

Since the second derivative is not suitable for random noise, the relationship between the mean (m) and variance (S^2) of soil volumetric water content within 48 hours before and after time t is added as a detection condition (Equation 8). If the observation value simultaneously satisfies Equations (5), (6), and (8), it is considered an abnormal peak.

For example, at Hubei Jingshan Station during July 2019, the first-layer soil volumetric water content was abnormal (marked in red in Figure 3c), remaining at 39.8% for an extended period, which is inconsistent with soil moisture variation characteristics.

2.3.5 Constant Value Detection (CD)

Constant value detection, also called minimum variation detection, addresses unrealistic observation records caused by instrument failure, frost, or other factors resulting in unchanged or minimally changed values over time. To distinguish normal oversaturation caused by precipitation, the condition must persist for a specified duration with soil volumetric water content variation smaller than sensor precision.

For 10–20 cm depth: $\max(48\,\mathrm{h}) - \min(48\,\mathrm{h}) < 0.0005$
For 30–50 cm depth: $\max(15\,\mathrm{d}) - \min(15\,\mathrm{d}) < 0.0005$

For example, at Guangdong Zijin Station during July 2019, the first-layer soil volumetric water content was abnormal (marked in red in Figure 3d), showing no obvious positive response to precipitation and fluctuating between 26.5%–36.5%, which is inconsistent with soil moisture variation characteristics.

2.4 Detection Process

Following the spectral analysis approach proposed by Dorigo et al. [6], we established a soil moisture data quality control and comprehensive analysis process. Referencing surface meteorological observation data quality control flags [13], we designed quality flags for soil moisture observation data. The abnormal data detection process is shown in Figure 4.

If data are missing or not reported, quality control codes for all elements in the missing layer are marked as 2, and no further checks are performed. If frequency checks are not passed, the sensor is deemed faulty for that layer, all elements are judged erroneous, and quality control codes are marked as 2, with no further checks performed. If volumetric water content and relative humidity threshold checks are not passed, quality control codes are marked as 1, and no further checks are performed.

Among the three checks—no-precipitation sudden change detection, abnormal peak detection, and constant value detection—if any check is not passed, the data are judged erroneous or suspicious, and the volumetric water content quality is marked as 1. If all three checks are passed, the quality is marked as 0. Other elements use the same quality control codes as volumetric water content.

3.1 Sensor Failure Detection Based on Frequency Characteristics

Using data from 2,012 automatic soil moisture stations nationwide in 2015, we analyzed the distribution characteristics of sensor frequencies in air and water (Figure 5). Water frequency values were mainly concentrated between 67–72 MHz, with 71.90% of sensors in water falling within the 67–72 MHz range. Air frequency values were mainly concentrated between 41–44 MHz, with 72.31% of sensors in air falling within the 41–44 MHz range. Due to differences in soil organic matter content and pH across regions, frequency boundaries vary slightly.

Table 1 compares the abnormal value detection results between frequency detection and volumetric water content detection for 101 automatic soil moisture stations in Xinjiang Uygur Autonomous Region during June–September 2015. The FD method detected 1,248 abnormal values, all confirmed to be caused by sensor failure at stations including Fuhai, demonstrating that FD can better locate the source of data quality problems.

The spatial distribution of abnormal values detected by the two methods shows that FD-detected abnormal values are more numerous in northern and central Xinjiang, while western Xinjiang stations generally have fewer abnormal values. Therefore, stations with higher FD detection rates generally have lower effective data quality, while those with fewer FD-detected abnormal values have higher quality. VD-detected abnormal values are generally fewer than FD-detected ones, except at Paotai Station where VD detection reached 1,248 abnormal values.

Figure 6 shows the spatial distribution of abnormal values detected by frequency and volumetric water content in Xinjiang Uygur Autonomous Region in 2015. The spatial distribution of abnormal data quality detected by the three methods is presented in Figure 7. FD-detected abnormal values are relatively high, primarily characterized by long-term quality problems at a small number of stations. Among these, SBD-detected abnormal values reach up to 48×10^5, accounting for 10.5% of the network, with Qinghai showing the most prominent issues. This type of quality problem is mainly caused by inaccurate manual determination of soil hydrophysical constants. Yuepuhu also shows relatively high SBD-detected abnormal values.

3.2 Abnormal Value Verification Analysis Based on Volumetric Water Content

Applying the abnormal data detection scheme and process described in Section 2.4 to hourly data from 2,012 automatic soil moisture stations nationwide, the data accuracy rate is 93.8%, the suspicious/erroneous rate is 1.7%, and the missing rate is 4.5%. Abnormal data are mainly distributed in Qinghai, Sichuan, Shandong, Hebei, and Guangdong provinces. The soil surface layer (0–20 cm) has a higher suspicious/erroneous rate than other layers, and only a few stations exhibit anomalies across the entire observation layer. Most automatic soil moisture stations nationwide exhibit data quality anomalies, primarily characterized by long-term quality problems at a small number of stations caused by sensor calibration or inaccurate determination of soil hydrophysical constants.

Figure 8 shows the vertical distribution of abnormal data across layers. Soil moisture abnormal values are mainly concentrated in shallow and middle layers (0–50 cm). The number of stations detecting abnormal values is relatively stable across layers, fluctuating between 80–126, but the suspicious/erroneous rate shows a trend of first increasing, then decreasing, then increasing again with observation depth. The SBD suspicious/erroneous rate peaks at 0.44%, mainly occurring in Sichuan and Guangdong provinces, possibly related to underlying surface conditions or groundwater levels (Figure 8a). The number of stations detecting abnormal values shows a decreasing trend with depth, with large fluctuations ranging 40–146, while the suspicious/erroneous rate remains stable, basically fluctuating between 1.28%–1.43% (Figure 8b). The PD suspicious/erroneous rate shows a clear decreasing trend with observation depth, peaking at 0.19% in the surface 10 cm layer, which is easily affected by external environments (precipitation, evaporation, etc.) (Figure 8c).

Figure 9 shows the spatial distribution of stations with abnormal values detected across the entire observation layer by the three methods. Spatially, these are mainly concentrated in Xinjiang, Qinghai, Sichuan, Henan, and Guizhou, alerting equipment maintenance personnel to perform timely calibration and maintenance.

4 Conclusions

To eliminate abnormal data from automatic soil moisture observation records, this paper proposes a set of anomaly detection methods applicable to automatic soil moisture observations. Using historical data from automatic soil moisture stations, we investigated the inherent variation characteristics of automatic soil moisture observations. Combined with instrument observation principles, data characteristics, and error sources of abnormal data, we designed five detection methods: FD, VD & RHD, SBD, PD, and CD. Using data from June–September 2019 from national automatic soil moisture observation stations, we verified the application effect of these detection methods. The results indicate:

1) The frequency-based instrument failure detection method, by introducing variation characteristics of sensor direct measurement values in air and water, can effectively detect erroneous data caused by instrument failure, with an accuracy rate reaching 100%.

2) Addressing prominent quality problems in soil moisture observation, the four detection methods based on inherent characteristics of soil moisture observation data and error sources of abnormal data can effectively detect anomalies in soil moisture observations, helping users quickly identify and resolve quality problems at the observation end.

3) The evaluation results from hourly data of national automatic soil moisture stations from June–September 2019 show a data accuracy rate of 93.8%, suspicious/erroneous rate of 1.7%, and missing rate of 4.5%. Abnormal data are mainly distributed in Qinghai, Sichuan, Shandong, Hebei, and Guangdong provinces. The soil surface layer (0–20 cm) shows a higher suspicious/erroneous rate than other layers, with only a few stations exhibiting anomalies across the entire observation layer. Although most national automatic soil moisture stations exhibit data quality anomalies, the primary issue involves long-term quality problems at a small number of stations, mainly caused by sensor calibration or inaccurate determination of soil hydrophysical constants.

4) This method has been preliminarily applied to the national Integrated Meteorology Observation Data Quality Control Business System (Tianheng). To further improve the applicability and accuracy of this method, future research should focus on automatic detection methods for various abnormal data in actual soil moisture observations, particularly persistent

abnormal peaks and sudden drops, to further enhance soil moisture data quality.

References

[1] Zhang Zhifu. Quality control schemes for hourly soil moisture from automatic weather stations in China[J]. Arid Land Geography, 2013, 36(1): 101-108.

[2] Wang Xiaodong, Yang Taiming, Wu Biwen, et al. Error analysis and quality control method of automatic soil moisture measurements in Anhui[J]. Meteorological Science and Technology, 2015, 43(3): 399-404.

[3] Wang Liangyu, He Yanbo, Zhang Yanhong, et al. Data quality early warning of automatic soil water observations based on soil hydrological constants[J]. Meteorological Science and Technology, 2016, 44(4): 528-534.

[4] Wu Dongli, Cao Tingting, Xue Hongxi. The study of quality control for observing data of automatic soil moisture[J]. Hans Journal of Soil Science, 2016, 4(1): 1-10.

[5] Wang Jiaqiang, Zhao Yufei, Ren Zhihua, et al. Design and verification of quality control methods for automatic soil moisture observation data in China[J]. Meteorological Monthly, 2018, 44(2): 244-257.

[6] Dorigo W A, Xaver A, Vreugdenhil M, et al. Global automated quality control of in situ soil moisture data from the International Soil Moisture Network[J]. Vadose Zone Journal, 2013, 12(3): 918-924.

[7] Wu Qinqin, Mo Shuhong, Cheng Shengdong, et al. Study on the correlation between soil moisture and temperature of different land uses in the loess area during a freezing thawing period[J]. Arid Zone Research, 2020, 37(3): 627-635.

[8] Zhao Chunlei, Shao Ming'an, Jia Xiaoxu, et al. Spatial distribution of water in the active soil layer along the south-north transect in the Loess Plateau of China[J]. Journal of Arid Land, 2019, 11(2): 228-240.

[9] Gong Yidan, Xing Xuguang, Wang Weihua. Factors determining soil water heterogeneity on the Chinese Loess Plateau as based on an empirical mode decomposition method[J]. Journal of Arid Land, 2020, 12(3): 462-472.

[10] China Meteorological Administration. Specification for automatic soil moisture observation (trial)[R]. Beijing: China Meteorological Administration, 2010.

[11] Basara J B, Crawford T M. Improved installation procedures for layer soil moisture measurements[J]. Journal of Atmospheric and Oceanic Technology, 2000, 17(6): 879-884.

[12] You J S, Hubbard K G, Mahmood R, et al. Quality control of soil water data in applied climate information system: Case study in Nebraska[J]. Journal of Hydrologic Engineering, 2010, 15: 200-209.

[13] China Meteorological Administration. QXT118-2010. Quality control of surface meteorological observational data[S]. Beijing: China Meteorological Press, 2010.

[14] Wang Liangyu, He Yanbo. Research on outlier threshold of automatic soil moisture observation data[J]. Meteorological Monthly, 2015, 41(8): 1017-1022.

[15] Hubbard K G, Goddard S, Sorensen W D, et al. Performance of quality assurance procedures for an applied climate information system[J]. Journal of Atmospheric and Oceanic Technology, 2005, 22: 105-112.

[16] Gonzálezrouco J F, Jiménez J L, Quesada V, et al. Quality control and homogeneity of precipitation data in the southwest of Europe[J]. Journal of Climate, 2001, 14(14): 964-978.

[17] Journée M, Bertrand C. Quality control of solar radiation data within the RMIB solar measurements network[J]. Solar Energy, 2011, 85: 72-86.

[18] Pei Buxiang, Mao Fei. Study on soil moisture variation patterns and the rational distribution of soil moisture observation stations of China[J]. Acta Agriculture Universities Pekinensis, 1990, 16(Suppl. 1): 1-12.

[19] Cai Lianghong, Ding Jianli. Remote sensing inversion of soil moisture based on modified vegetation index model analysis and comparison[J]. Arid Land Geography, 2017, 40(6): 1248-1255.

[20] Ding Xinyuan, Zhou Zhibin, Lei Jiaqiang, et al. Analysis and comparison of models for soil water characteristic curves of Tarim Desert Highway Shelterbelt[J]. Arid Land Geography, 2015, 38(5): 985-993.

[21] Chen Bingyin, Yang Liao, Chen Xi, et al. Application of modified TVDI in drought monitoring in arid areas[J]. Arid Land Geography, 2019, 42(4): 902-913.

[22] Wang Xinyou, Ma Quanlin, Jin Hujia, et al. Soil respiration variation characteristics and its relationship with hydrothermic factor of artificial Haloxylon ammodendron forest in lower reaches of Shiyang River[J]. Arid Land Geography, 2019, 42(3): 570-580.

[23] Wang Jun, Li Guang, Nie Zhigang, et al. Testing a soil water erosion predictive model in an artificial grassland in the hill gully Loess Plateau[J]. Arid Land Geography, 2020, 43(2): 398-405.

[24] Han Jiaojiao, Duan Xu, Zhao Yangyi. Spatial and temporal variability of soil moisture on slope land of different vegetation of dry hot valley in Jinsha River[J]. Arid Land Geography, 2019, 42(1): 121-129.

[25] Lü Mingxia, Wang Yibo, Liu Guohua. Changes in shallow soil moisture and its influencing factors in the alpine region of the upper Heihe River[J]. Arid Zone Research, 2020, 37(4): 899-907.

Note: Figure translations are in progress. See original paper for figures.