

Curating 62 years of Walnut Gulch Experimental Watershed dataset: Improving QAQC methods of rainfall and runoff measurements.

Menberu Bitew¹, Eleonora M. C. Demaria², Philip Heilman¹, David C. Goodrich², Mark Kautz¹, Gerardo Armendariz¹, Carl Unkrich¹, Haiyan Wei³, and Anandraj Perumal³

¹USDA Agricultural Research Services

²USDA-ARS

³University of Arizona

October 1, 2020

Abstract

The Walnut Gulch Experimental Watershed (WGEW) is the primary outdoor hydrologic laboratory for the USDA-ARS' Southwest Watershed Research Center (SWRC). This site represents the Southwest semiarid environment within the Long-term Agroecosystem Research (LTAR) network. The SWRC maintains a collection of long-term hydro-climatic measurements from WGEW, featuring an extensive archive of rainfall and runoff observations from an ephemeral network of streams within the 149 km² watershed. The WGEW was established in 1953, and has continually developed and improved quality assurance and quality control procedures to aid in the accuracy and curation of the constantly growing datasets obtained from more than 100 rain gauges and 18 flumes, weirs, and gauged ponds. These efforts have led to the development of a state-of-the-art database and data visualization tools to aid in the curation of research-grade hydrometeorologic datasets. This required development of automated quality assurance and quality control tools to check and maintain the data for 21st century research needs. We developed five tools to improve the quality of rainfall and runoff database based on conventional hydrologic principles and the relationships between them: 1) precipitation is spatially correlated; 2) there is a temporal relation between rainfall and runoff; 3) runoff is only a limit portion of rainfall; and, 4) closer took of extreme events. Hence, we developed the following methods that included the analysis of interpolated rainfall maps at a daily time step, the association between rainfall and runoff events, lag time, runoff coefficients, and multiple regression methods to identify problematic events in the data archive. To visually inspect and verify the errors, we developed a graphical tool that displays relevant event hyetographs and hydrographs within a specific time window. After flagging anomalous events, we evaluated the types of errors using the original records and metadata information. The implementation of these approaches resulted in developing a suite of semi-automated QAQC tools that correctly detected 813 rainfall and 24 runoff events with erroneous timestamps that had passed all previous quality checks.

Curating 62 years of Walnut Gulch Experimental Watershed dataset: Improving QAQC methods of rainfall and runoff measurements.

Menberu B. Meles^{1,2}, Eleonora M.C. Demaria¹, Philip Heilman¹, David C. Goodrich¹, Mark A. Kautz¹, Gerardo Armendariz¹, Carl Unkrich¹, Haiyan Wei², Anandraj Thiagaraja Perumal

¹USDA-ARS, Southwest Watershed Research Center, Tucson, AZ, USA.

² University of Arizona, School of Natural Resources and the Environment, Tucson, AZ, U.S.A

Corresponding author: Menberu B. Meles (mbmeles@email.arizona.edu)

Key Points:

- Long-term rainfall and runoff observations require continual quality checking and curation to improve fidelity.
- Evaluating the relationships of conventional hydrologic principles plays a paramount role in the development of new QAQC tools to flag questionable rainfall and runoff events.
- Quality hydrologic data reduces the uncertainty in hydrologic analyses and model predictions.

Abstract

The Walnut Gulch Experimental Watershed (WGEW) is the primary outdoor hydrologic laboratory for the USDA-ARS' Southwest Watershed Research Center (SWRC). This site represents the semiarid environment of the Southwest U.S. within the Long-term Agroecosystem Research (LTAR) network. The SWRC maintains a collection of long-term hydro-climatic measurements from WGEW, featuring an extensive archive of rainfall and runoff observations from an ephemeral network of streams within the 149 km² watershed. The WGEW was established in 1953, and has continually developed and improved quality assurance and quality control (QAQC) procedures to aid in the accuracy and curation of the constantly growing datasets obtained from more than 100 rain gauges and 18 flumes, weirs, and gauged ponds. These efforts have led to the development of a state-of-the-art database and data visualization tools to aid in the curation of research grade hydrometeorologic datasets. This required development of automated quality assurance and quality control (QAQC) tools to check and maintain the data for 21st century research needs. To improve the integrity of the historical rainfall and runoff database, five independent tools were developed to identify questionable events based on the following conventional hydrologic principles and the relationships between them: 1) precipitation is spatially correlated; 2) there is a temporal relation between rainfall and runoff; 3) there is an upper limit to the amount of runoff generated from a given rainfall event; and, 4) it is always important to verify extreme events. Hence, we developed and applied the following methods that included the analysis of interpolated rainfall maps at a daily time step, the association between paired rainfall and runoff events, the computation of runoff lag time, runoff coefficients, and the analysis of multiple regression methods to identify problematic events in the data archive. To visually inspect and verify the errors, we developed a graphical tool that displays relevant event hyetographs and hydrographs within a specific window of time. After flagging anomalous events, we evaluated the types of errors using the information in the original records and metadata. The implementation of these approaches resulted in the development of a suite of semi-automated QAQC tools that correctly detected 813 rainfall and 24 runoff events with erroneous timestamps that had passed all previous quality checks.

Keywords: QAQC, hydrologic database, long-term data curation, rainfall and runoff measurements, Walnut Gulch, Quality hydrologic data

1. Introduction

Long-term observations are the foundation for understanding earth system processes for sustainable water and natural resource management under current and future climate and management conditions (*Tetzlaff et al., 2017*). As the uncertainty in water supply increases with warmer temperatures and growing human population, the need for quality, long-term hydrometeorological observations has increased. Data collected in experimental watersheds and natural research observatories across the U.S., such as the USDA-ARS Long-Term Agroecosystem Research (LTAR) network (*Kleinman et al., 2019; Spiegel et al., 2018; Walbridge and Shafer, 2011*), the Critical Zone Observatories (CZO) (*Brantley et al., 2017*), the Long-term Ecological Research (LTER) network, (*Knapp et al., 2012*), and the National Ecological Observatory Network (NEON) sites (*Collinge, 2018*) are critical for research, analyses, and planning (*Tetzlaff et al., 2017*). Ensuring that the accuracy of these measurements are suitable for scientific analysis and long-term monitoring requires a large investment in time and money. Furthermore, these efforts need to be operationally feasible and address the biophysical diversity of the locations within these research networks. Common methods to ensure the integrity of these data, referred to as quality assurance and quality control (QAQC) procedures, include outlier checks using thresholds, comparisons with neighboring stations, and fitting to theoretical probability

distribution functions to flag outliers (*Kunkel et al., 2005; You et al., 2007*). QAQC procedures reflect improvements to the quality of a dataset through a combination of steps that prevent errors from occurring, quality assurance (QA), through improved data collection and management practices, and the identification and resolution of errors within the recorded dataset, quality control (QC).

Such efforts have been under continuous development at the Walnut Gulch Experimental Watershed (WGEW) in southern Arizona. This site is operated by the USDA Agricultural Research Service (ARS) Southwest Watershed Research Center (SWRC) and is an LTAR site representing the semiarid grazing lands of the southwestern United States. The SWRC has invested millions of dollars since the establishment of WGEW in both QA, through regular sensor calibrations, site maintenance, installation of consistent instrumentation across networks, thorough testing of new instrumentation prior to adoption, and redundant sensors to compare observations (e.g., *Fiebrich and Crawford, 2001*), and QC through regular data inspection and analysis conducted by SWRC scientists and technicians.

In semiarid regions in the Southwestern U.S., where natural and human systems compete for already over-allocated water resources, accurate hydrologic measurements are essential. The WGEW was established during the mid-1950s (*Renard et al., 1993; Moran et al., 2008a*) to measure high resolution rainfall, runoff and other hydro-climatic variables within in a semiarid, ephemeral network of streams. QAQC procedures at WGEW have evolved to accommodate the observations made by retired analog instrumentation (*Brakensiek et al., 1979*) and current digital instrumentation (*Moran et al., 2008a*). The long-term record of measurements from the WGEW are archived in the Data Access Project (DAP - <https://www.tucson.ars.ag.gov/dap/>) database which was initiated in the year 2000 when analog instrumentation was replaced with digital instrumentation (*Nichols and Anson, 2008*). Rainfall and runoff observations in the WGEW are routinely checked for inconsistencies and errors such as problematic recession curves on runoff hydrographs or timing errors in measurement. Despite the past QAQC efforts, several errors including human mistakes and instrument malfunctions remain in the data, largely due to time stamp errors associated with analog clocks that were used prior to 2000. QC efforts have historically been performed on rainfall and runoff observations independently, however, given the causality between these measurements, a new QAQC procedure was needed that considered their relationships in order to identify inconsistencies in the data and further improve the overall quality of DAP.

From an ecosystem analysis standpoint, errors in rainfall or runoff observations could lead to inconsistencies in model parameterization, in model predictions, and errors in water balance partitioning. In this paper, we initially used watershed-wide interpolated rainfall maps to isolate observations that are inconsistent in space and time, which is expanded to include the relations between the long-term data records to execute a series of spatial and temporal comparisons to evaluate data inconsistency. The goals of this study were: 1) to design and implement systematic methods for identifying erroneous rainfall and runoff data in DAP using relationships between the two datasets based on conventional hydrologic principles for semiarid regions; and, 2) to identify, verify, and remove the inconsistent rainfall and runoff observations in the DAP database.

2. Data description and quality control methodology

2.1. WGEW data and site description

The study was conducted in three sub-watersheds at WGEW, WS-04, WS-011, and WS-06 (Figure 1a) with 2.3, 8.2, and 95.1 km² contributing areas respectively. In each of the sub-watersheds, there are 5, 12, and 56 rain gauges and a runoff recording flume at each outlet. Average summer rainfall, defined as the rainfall observed in the period May through September, over the watershed for the 1961-2017 period has been measured at approximately 293 mm (112 to 370 mm range), which represents approximately 60% of the annual total (*Demaria et al., 2019*). Summer rainfall is important in this environment, as almost 100% of the runoff is generated during this period. To accurately measure the rainfall from spatially isolated thunderstorms and the resulting runoff during the summer season, WGEW is equipped with a dense network of rain gauges and runoff measurement structures. Based on experiences from eastern and mid-western watersheds, 20 rain gauges were initially installed across the WGEW which increased to 99 rain gauges over time to capture

the high spatiotemporal variability of rainfall generated from highly localized air-mass convective storms during the summer monsoon season. (Goodrich *et al.*, 1997). Rain gauge density per area is about one gauge per 1.7 km², making the WGEW the densest rainfall network on a semiarid experimental watershed in the world (Garcia *et al.*, 2008; Goodrich *et al.* 2008a;). Since the establishment of ARS experimental watersheds and ranges in other regions of the country starting in the mid-1930s, the primary precipitation gauges employed have been a Belfort 8-inch (0.2032 m) unshielded weighing-bucket gauge (Figure 1b). If well calibrated, these gauges have a measuring resolution of 0.01 inches (0.254 mm). Runoff response has been recorded at 18 nested sub-watersheds at WGEW using v-notch weirs, H flumes, and supercritical flumes designed for Walnut Gulch channel environment (Stone *et al.*, 2008; Smith *et al.* 1982) (see. Figure 1c for supercritical flume photo). The WGEW watershed and its surrounding area are equipped with a total of 132 rain gauges, 18 runoff measurement flumes and weirs, three meteorological stations, two eddy covariance flux towers, soil moisture sensors at various depths, and eight gauged tanks/ponds. The temporal frequency of the recordings varies based on the type of sensor, with rainfall and runoff observations being recorded at a maximum frequency of one minute. In addition to the observations representing hydrology, climate, and sediment, the SWRC has extensively collected datasets related to vegetation abundance and diversity and conducted extensive research in geology, soils, hydrology, ecology, and rangeland studies which provide a broad set of data and proxies to investigate changes in water resources of this semiarid environment. Common instrumentation monitoring, calibration, and maintenance procedures have been followed through time (Brakensiek *et al.*, 1979). All WGEW data (current and historical) are available through the web-based portal for the DAP database (<https://www.tucson.ars.ag.gov/dap/>).

With evolving digital technology, in the mid-1990s, the mechanical chart recorders on instruments were replaced with electromechanical components and digital recorders, eliminating the labor-intensive weekly rain gauges visits, often over rough roads, to wind clocks, add ink to recording pens, and change charts. After field collection of the paper charts, they must be manually digitized to convert the pen trace to breakpoint digital format (Goodrich *et al.*, 2008). Both rainfall and runoff observations are archived at sub-hourly frequency using breakpoint format, which consists the times of change from one steady rate to another with a minimum time step of 1 minute. In the breakpoint data, the stream of data pairs consists of time and rainfall or runoff depth with zero values at the beginning and end of the event recordings. A side-by-side direct comparison of events before and after the change from analog to digital gauges using pairs of co-located analog/digital rain gauges was conducted for a 5year period for nine (9) 24 hour gauges from 2000-2004. This analysis concluded that no measurement discontinuities of a number of raw and derived quantities (e.g. peak 30 min rainfall intensity) were introduced by the switch in the weighing and recording systems (Keefer *et al.*, 2008). In general, we observed that the analog period produced more errors than the digital period, which could be attributed to mechanical clocks slowing down, speeding up, or stopping between visits as well as expansion or contraction of paper charts, and the manual digitizing process.

The DAP database has been used extensively in several hydro-ecological modeling studies (Costa *et al.*, 2012; Duan *et al.*, 2006; Goodrich *et al.*, 2004; Goodrich *et al.*, 2012; Niu *et al.*, 2014; Scott and Biederman, 2017) and in remote sensing studies as a critical ground validation site for retrieval algorithms, including soil moisture and vegetation measuring platforms (Morin *et al.*, 2003; Moran *et al.*, 2008b; Amitai *et al.*, 2012; Das *et al.*, 2012; Knipper *et al.*, 2017; Kolassa *et al.*, 2018). Therefore, the DAP observational database must be as free of errors as possible. A special section of 20 papers in Water Resources Research provides WGEW history, data collection, metadata, data access and recent results of WGEW data analysis (Moran *et al.*, 2008a – all the papers in the special section can be found at <https://www.tucson.ars.ag.gov/dap/wrr.html>).

2.2. Existing QAQC processes

The objective of the existing QAQC procedure is to prevent erroneous information from getting into the database and to facilitate maintenance of equipment efficiently. Substantial effort has been invested over the years to ensure the quality of the WGEW rainfall and runoff datasets. Rainfall depth and runoff stage measurements are calibrated and adjusted annually prior to the monsoon season. Processing workflow from field measurement to storage in the database and public access through a web-interface requires several steps

to ensure consistent quality. The raw data and its summary report are transmitted to a central location daily and are internally accessible for monitoring, identifying malfunctions, and deployment of maintenance technicians. The data then undergoes a QC processes including visual inspection of the hyetographs and hydrographs, and geospatial representation of rain gauge locations with associated rainfall amounts at the field office for approval every 2-3 weeks. Using these procedures, a significant portion of the errors are removed, however, given the difficulty of identifying numerous electromechanical errors or malfunctions, environmental issues (e.g. bird nests in gauge orifice), technician errors, and operational changes, maintaining quality data acquisition and storage faces significant challenges. There have been several instances in which the existing approach has not detected errors such as inaccurate base flow values over a long period of time, incorrect time stamps of the recordings, and falsely generated data from electromechanical failure of sensors. Moreover, the analog system in use until 2000 required a labor-intensive workflow resulting in increased chance of human errors and the inability to identify anomalous recordings for almost five decades worth of data. Within the 2 million gauge-day records, some errors are to be expected. Errors might have occurred due to the changes in the equipment used for data reading, storage, and digitizing recorded charts. The current challenge is to find new ways to improve data quality by identifying and correcting errors within the extensive database (*Bitew, et al., 2019*). In addition to the potential errors listed, there were other problems such as decommissioning a large number of rain gauges between 1982 to 1989 during the winter season (DAP provides complete metadata for these dates), when frontal rainfall dominates and is more spatially uniform, due to budgetary constraints. In this period, only nine gauges on the watershed were operated year-round (*Nichols et al., 2002*).

2.3. Methods

Fundamental hydrologic principles and relationships were used to develop the new QAQC tools for auto-flagging the inconsistencies and errors in the WGEW data. These are:

- 1) Precipitation is spatially correlated;
- 2) Precipitation can be linked temporally to runoff;
- 3) There is a limit to the amount of precipitation converted to runoff; and, 4) Extremes events require additional examination.

Spatial correlation of rainfall: Even though the summer rain in WGEW is characterized by localized and short-lived events (*Goodrich et al., 1997*), the dense rain gauge network in the watershed is assumed to capture the spatial structure of the rainfall at the event scale (*Osborn et al., 1980*). Rainfall observations from rain gauges close to one another show strong spatial correlation, which weakens as the separation distance increases. *Keefer et al. (2015)* identified correlation decreases significantly as the spatial distance increases more than 2 km between rain gauges using 59 rain gauges 1961–2013 in WGEW. This principle necessitates the application of interpolation techniques in the context of identifying erroneous events.

Correlation of rainfall and runoff: There is a temporal correlation between rainfall and runoff observations. Based on this, two relationships were utilized: 1) a one-to-one association of rainfall and runoff (Association Method) by searching for the amount of rain observed within a given window of time for each runoff observations, and 2) computing the value of delay time between the runoff and the time of precipitation occurrence (Lag Time Method). For each runoff event, there exists a corresponding rainfall event that caused the runoff, and the delay time should not be too early or too late.

Runoff is a small portion of precipitation: Not all rainfall is converted to runoff. All or a portion of the rainfall is usually abstracted in the landscape resulting in small runoff to rainfall ratio. The abstraction is usually large in arid and semiarid regions (*Goodrich et al., 1997*) resulting in only a small portion of the rainfall converted into the runoff. The runoff amount is the result of complex integrated hydrologic functions based on the rainfall and watershed properties. These include rainfall attributes (spatial extent, depth, intensity, duration, etc.) and the watershed properties, including ephemeral channels, ground cover, soil, slope shape, etc. This principle was applied in two forms: To evaluate the temporal relationship between

runoff and precipitation and watershed properties and antecedent condition (Regression Method; *Bitew et al., 2019*); and, To calculate an index that accounts for the ratio of runoff to rainfall amount (Runoff Coefficient Method).

2.3.1. Rainfall Interpolation method

To improve process understanding and produce visually intuitive gridded maps of rainfall, we performed a 100 m X 100 m gridded spatial interpolation across the entire WGEW rainfall dataset for the period 1954-2014 at the daily scale. The interpolation method is based on a Multiquadric-Biharmonic (MQ-B) interpolation, as described in *Garcia et al. (2008)* and *Syed et al., (2002)* . By visually inspecting each map, we identified rain gauges with anomalously high/low rainfall amounts and flagged them. The interpolated maps were used as a first step to narrow the focus on potentially erroneous events.

The analysis of the interpolated rainfall maps identified two types of errors: 1) Type I Error: the result of rain gauges being turned off without any designation in the database or missing gauge data due to other reasons; and, 2) Type II Error: the result of date coding errors (i.e., the wrong date was recorded when digitizing the analog recording chart). For Type I Error cases, rainfall totals for a rain gauge are recorded as zero in the database due to equipment malfunction, human error, or the gauge being turned off without notation. Causes of Type II Errors include: an erroneous date recorded for the rainfall (i.e., the rainfall was marked as occurring on a different date), false data produced from electromechanical failure or environmental interference, and human error in interpreting the analog data. For each gauge-rainfall day, we evaluated the map of interpolated observations to determine if inconsistencies on the rainfall field associated with the errors were verifiable or not. The date of each flagged rainfall event and the rain gauge number were stored for further verification and correction in the database (Figure 2). The first step in verifying the flagged interpolated daily rainfall inconsistencies is to identify if they belonged to the periods when the gauges were turned off for budgetary reasons. Those dates are usually available in the database metadata. In the case of the flagged dates outside the period when gauges were turned off, the field notes and paper charts were used in the verification processes.

2.3.2. Rainfall-runoff association method

The association method (Figure 3a) was designed based on the fact that in ephemeral channels, any runoff generation is always triggered by an associated precipitation event. If a given runoff event had no associated rainfall within a reasonable time window, it was flagged for further investigation. It was important that we considered timing errors for runoff events that did not have associated rainfall due to instrument time-keeping limitations during the period of analog recording. During this period there was no way to synchronize clocks other than to manually adjust to a wrist watch. Furthermore, attempts to correct data with known timing issues were often made by adjusting to the measurement of a nearby instrument.

In this method, we followed four steps to identify the erroneous events: 1) We generated continuous hourly rainfall and runoff data from the breakpoint data format in the database; 2) We then calculated the spatial sub-watershed average of the aggregated hourly rainfall using only those rain gauges within the sub-watershed boundary and only included the gauges that recorded rain, i.e., only non-zero values were included in the average. This approach was used to account for instances of relatively intense rain recorded in a small number of rain gauges within the sub-watershed that could have resulted in runoff, but if averaged over all the rain gauges could have resulted in a very small depth; 3) The hourly runoff data and the spatial hourly average rainfall data are then merged and summarized into total runoff volume within seven (7) hours windows and total rainfall depth within a 4-hr window; and, 4) Finally, we searched for a one-to-one association between the runoff volume and the corresponding rainfall depth, which resulted in flagging those events with insufficient (< 1 mm) or no recorded rainfall as problematic data. The event-based summary using the seven-hour window for the runoff and the four-hour window for rainfall were selected through trial and error across several sub-watersheds in WGEW. In summary, we accumulated the runoff depth within the seven-hour window; 3-hour preceding and 3-hour following the hour at which the maximum runoff was observed. The event-based rainfall depth was calculated as the sum of all rainfall accumulated 3-hours

proceeding and 1-hour succeeding the hour of maximum runoff observation for all rain gauges in for near the sub-watershed (Figure 3a).

2.3.3. Lag time method

The lag time method relies on the relationship between the temporal properties of the rainfall hyetographs and the corresponding runoff. The lag time is defined as the time interval between the center of mass of the hyetograph for a given rainfall event (CM_f) and the time of the peak runoff rate in the hydrograph for the associated runoff event (t_{peak}) (Figure 3b). The center of mass of excess rainfall (CM_f) was estimated as the median of excess rainfall time, i.e., rainfall duration after the starting time of the runoff (equation 1). In long-duration rainfall events where only rainfall in a certain portion of the duration has resulted in runoff, equation 1 may not provide representative CM_f value. In such cases, the use of equation 2 provides a better estimate of CM_f values. In most cases, the center of mass calculation based on the median of excess rainfall duration (equation 1) and time of maximum rainfall (equation 2) yielded similar results. Lag time was then calculated using equation 3. The lag time related parameters used in this analysis are the rainfall start time (t_{oprecp}), runoff start time ($t_{oRunoff}$), time of peak runoff rate (t_{peak}), time of maximum rainfall intensity ($t_{max.int}$), and rainfall duration (D) (Figure 3b).

In watersheds with multiple rain gauges, the hyetograph of the rain gauge with the minimum succeeding time difference i.e., $t_{oprecp} - t_{oRunoff}$ was used to calculate CM_f . Rainfall events with unrealistically long lag times could be due to errors in the rainfall or runoff data timestamps and were flagged.

$$CM_f = \frac{D - (t_{oprecp} - t_{oRunoff})}{2} \quad (1)$$

$$CM_f = t_{max.int} \quad (2)$$

$$Lag\ Time = CM_f - t_{peak} \quad (3)$$

The maximum threshold for lag times ranged from 1-2 hours and was defined for each sub-watershed based on the size of the drainage area. The estimated travel time, from the storm center to the outlet of the watershed, was determined by simple routing equations. Lag time was plotted against mean runoff rate for each event for visual identification of the possible thresholds. An example of the visual aid for threshold identification is given in Figure 4.

2.3.4 Runoff coefficient method

The runoff coefficient (C) is computed as the total event runoff volume divided by the total sub-watershed-averaged rainfall volume accumulated between the 3-hours preceding the beginning of runoff observations and the end of the runoff event (Figure 3c). The C values range from 0 to 1; 0 for no runoff generation from the rainfall and 1 when all rainfall over the sub-watershed is converted to runoff. In WGEW the runoff coefficient values range between roughly 0.065 at the hillslope scale and 0.006 over the entire watershed (Renard et al., 2008). This method was designed to identify two cases of rainfall-runoff relations: Case 1) C equal to infinity, which shows the condition of a runoff event without associated rainfall; and Case 2) an exceptionally large C approaching 1, which can raise concerns about the quality and accuracy of the rainfall and resulting runoff measurements. In Case 1, where C equals infinity, the error is likely related to the runoff data as in the case of the association method, but in Case 2 it is important to flag both runoff and rainfall for further investigation.

2.3.5. Regression method

We selected several hydrometeorological predictors from DAP with the potential to influence rainfall-runoff relations to develop a multi-parameter regression model (Figure 3d). The predictors that influence rainfall-runoff relationship in WGEW were evaluated using a multi-model inference approach that used the Akaike Information Criteria (AIC; Akaike 1973). In this approach, we developed a regression equation using the 18 predictors that describe rainfall properties, watershed properties, and the antecedent soil moisture conditions

related to the contributing area and channel system. The simulated runoff depth predicted by the regression model was then compared with the measured runoff to identify events with questionable rainfall-runoff relations based on the following three conditions. The first condition was the identification of substantial rainfall events (>20 mm depth) that did not produce runoff. The second condition identified rainfall events with observed runoff (>0.5 mm depth), but zero predicted runoff. The third condition identified those rainfall events with observed runoff, but no predicted flow. For details of the development and performance of the regression model in QAQC application in the WGEW case, see *Bitew et al. (2019)*.

2.4. Verification/Validation

Once the questionable rainfall and runoff events were flagged using the test methods described above, the flagged events were further investigated by visual inspection using a plotting tool that displays the runoff event and corresponding rainfall. The plotting tool displays cumulative runoff volume, the runoff rate, the corresponding cumulative rainfall as the spatial mean of the watershed, and the cumulative rainfall from each of the rain gauges in the sub-watershed (see Figure 3e). In addition to using the plots for verification of the flagged events, historical metadata were also used to gain insight to the cause of the error. These data included: the original field notes and maintenance logs kept by technicians, note on the original paper charts from the analog instrumentation, and the use of information available from the neighboring sensors.

3. Results and Discussion

3.1 Rainfall observation errors

Using the rainfall interpolation method, two types of errors were identified. The Type I error is visually observed as isolated points with no measured rainfall surrounded by gauges reporting rainfall on the maps (Figure 5a). Type II errors (Figure 5b) can be identified on the rainfall maps when a single gauge reports rain or suspicious subset of gauges report rainfall while no other rainfall is detected in the watershed. Often times these gauges show Type 1 error on previous or subsequent days. Another relatively common error in the analog era was the clock stopping. If ink was still in the recording pen, partial data could sometimes be salvaged. In these cases, an estimated hyetograph was constructed using the temporal storm distribution of adjacent gauges. These instances are documented in the metadata for the gauge.

For each day with observed rainfall, a total of 6004 days in the 62-year period, we visually inspected each map and identified the number of occurrences of each type of error. Figure 6 a and b show temporal changes in error type as cumulative relative frequency (%) for daily rainfall during the period 1954-2014. Type I Errors are primarily associated with rain gauges not being operational for intentional and unintended reasons. From the total number of cases, $n = 641$, 28% occurred during the summer months (July -September) while 72% were observed during the winter months (October – March). The number of occurrences of Type I errors was largest in the 1980s and mainly during the winter months, a period when rain gauges were turned off due to budgetary restrictions. The percentage of reported Type I errors increased from 34% in 1978 to 76% in 1988. Type II errors totaled 160 cases during the 62-yr study period and were uniformly distributed through time except in 1962 and 1970 that showed 17 and 25 cases, respectively. For both error types, the majority of cases were reported during the analog recording period (pre-2000) indicating that despite the technical expertise of the WGEW technicians and the QAQC in place, human and mechanical equipment errors are sometimes difficult to detect and avoid. Of all the Type I errors, 49% of the cases were uncorrectable (unknown sources or lost information, etc.) while 41% were correctable cases (information available in the original charts or maintenance log) (see Figure 6 c and d). For Type II errors, the largest contribution to total error was correctable (56%) while the false alarm cases (33%) were not errors but identified as errors due to visual limitations in inspection of the maps.

3.2. Runoff observation errors

We applied the four QAQC runoff tests to all runoff-producing events on the 3 sub-watersheds. The analysis found a total of 102 questionable events flagged by one or more of the tests: 40 events in WS-04 (12% of events), 28 events in WS-11 (6% of events), and 40 events in WS-06 (6% of events). The tests identified

runoff events that were assigned erroneous time stamps that occurred due to malfunctions in the analog clocks, and human error when digitizing the charts. The errors in the runoff time stamps resulted, in some instances, as runoff occurring before the onset of the rainfall, in other instances it resulted from excessively large lag times or in excess of the threshold time used in the association and runoff coefficient methods. We also identified runoff events with excessively long recession curves that commonly occur in WGEW due to the functionality of the measurement flumes. These long, low-flow periods, initiated at the recession of the hydrograph, are a result of sediment deposition within the stilling well of the flume. This accumulation of sediment slows the drainage from the stilling well and consequently affects the measurement of the water level made by a floatation mechanism. These erroneous events are typically adjusted in the existing QAQC protocol by fitting a theoretical curve to the end of the recession in the hydrograph and noting in the metadata. Unfortunately, some of these types of errors have slipped into the database over the years. These errors were easy to identify with the association method due to the excessively long period of artificial low flow following the recession of the hydrograph. The association method identified these events because the 7-hr window breaks the long runoff duration data into multiple 7-hr runoff events where only the first window will have associated rainfall.

Overall, there were few instances where the four methods simultaneously flagged runoff events. The Lag time method flagged the largest percentage of events ranging from 51 to 80% for the 3 sub-watersheds, followed by the Regression method that identified between 40 and 50 % of the total number of events flagged (Figure 7 a). The Runoff Coefficient method flagged the smallest number of events and also showed the largest variability between the watersheds. We did not find a direct relationship between the efficiency of any of the four methods and the area of the sub-watershed. Each method, regardless of which one is used, can identify between 50-75% of the total number of flagged runoff events, but that percentage decreases to 7-26% for all the methods to detect the error simultaneously (Figure 7b).

The number of questionable events flagged depends on the threshold values selected in each of the methods. These thresholds require expert knowledge of the watershed being examined, and will vary depending on biophysical characteristics of the watershed. It is also important to note that the threshold values (7-hr window in the association method, 1-2-hr lag time, etc.) were determined iteratively based on detailed observation of the data analysis in this study. Examples of specific problematic events that were not previously identified as errors in the database were flagged by these tools and confirmed using the record information associated with the database (Figure 8). The figure includes a longer lag time than anticipated (Figure 8a and c), an excessive low flow duration (Figure 8b) that showed a false baseflow over more than 14 hours, and errors where runoff occurred before the rainfall (Figure 8e and f). Closer examination of the database and the inspection of analog charts for the questionable data points identified revealed different problems related to the timestamp of the data recording and storage. The events in Figure 8 (except b) were flagged using lag time, showing a significant delay of the rainfall that possibly resulted in the runoff. The association method flagged all events except e and f as all the runoff events did not have associated rainfall within the 4hr window leading to the maximum runoff hours. The main hydrograph of the event in Figure 8b (7/28/1966 on watershed WS11) does not seem to have a problem, but the fact that the runoff baseflow component in the recession side continued over 14 hours with the runoff rate value close to 0.2 cfs which appeared as a sequence of runoff events with no associated rainfall. The runoff coefficient method identified the events in a, c, and d, and the events in e and f were flagged by the regression method (see more on *Bitew et al., 2019*). We applied the plotting tool on about 90% of the 102 flagged events for which clock issues on rainfall/runoff were the reason for inconsistencies in the data. From the visual inspection, 72% of the flagged events were problematic while the rest required more than visual inspection to verify the error. The verified events had no rain corresponding to the runoff or the rainfall that appeared to have happened before the runoff event started. About 7% of the flagged events seem to be false alarms, which were identified mainly by the lag time and association method. The examples of the false alarms were dominantly runoff events created by long-duration variable rainfall across the watershed, which resulted in big lag time. There were instances in which runoff with multiple peaks or separate runoff events for the same long-duration rain occurred. In some cases, flow is recorded on the flumes after the contributing area is saturated sufficiently, resulting in

large lag time. In the event where the data showed that runoff recording before the rainfall in a few minute delay, the inspection was not very clear, which needed to compare the start time of the runoff and rainfall from all rain gauges in the watershed.

3.3. Strengths and weaknesses of the QAQC methods

Evaluations of the formulation of each of the methods and evaluating the list of flagged events have different strengths and weaknesses. For instance, in the interpolation method, the visual inspection of the daily interpolated rainfall maps clearly shows the expected spatial correlation between the rain gauges located close to each other. But this does not mean the maps that showed consistent spatial rainfall field are free of errors. The method misses errors related to sub-daily time steps. In the lag time method, the fact that start times, peak rate time, and duration are registered for all event rainfall and runoff makes it possible to evaluate the lag time at 1-minute time resolution. Using this method, we have identified errors that are very difficult to identify using other methods. Such errors include errors like when the runoff came 1-min before the start of the rainfall in the watershed. It is also important to note that runoff that appears to happen before the start of the rainfall could be related to runoff events triggered by heavy rain on the flumes. The representation of the center of mass of the rainfall and lag time computation using in the equations is a simplified version of the lag time. This representation, especially in long-duration rainfall and storms moving across the contributing areas, may result in erroneous lag time. The other limitation is a threshold lag time value, which usually depends on the watershed size. The accuracy of the arbitrary definition of the threshold lag time values is one of the limitations. The strengths and weaknesses of each of the methods are summarized in the Table 1 below.

4. Conclusions

Given the challenges facing natural resource management, every effort should be made to ensure that long-term datasets are used to their fullest potential. Implementation of carefully curated datasets, will ensure that future investigations will gain the benefits of reliable and trustworthy data. The methods used in this study were developed and applied to evaluate the quality of exceptional set of hydrometeorological data in a semi-arid environment that comprises the sub-hourly and high-density spatial observations. The use of relationships based on conventional hydrologic principles allows using the same principles in other regions. However, the application in different areas requires understanding of the unique features of rainfall-runoff responses in line with observation time steps and the observations' spatial density. Therefore, one can say this QAQC strategy can be potentially applied in similar networks with rainfall and runoff observations to improve the quality of their long-term databases with some modifications depending on specific meteorological and hydrological processes related to the watershed.

Prior to this study, there had not been a systematic linkage of rainfall to runoff for all of the nested sub-watersheds on WGEW. Inconsistencies in the database were successfully flagged due to mainly date/time code errors, events with unreasonably long low-flow periods, small runoff events generated with rainfall on flumes when it did not rain enough elsewhere in the watershed, maintenance flow mistakenly left in the database, etc. So far, the five tools were designed and applied in such a way to auto flag the problematic events separately. These tools can be combined into a single partially automated workflow process, but verification through visual inspection by a trained technician or scientist is still necessary. The agreement between the four methods was minimal, highlighting the importance of using more than one tool to flag possible errors in rainfall-runoff observations. Due to the assumptions in each of the methods, some errors are better identified by specific methods while they miss certain types of errors. Therefore, it is vital to apply all events and flag as many events as possible and go through one by one to confirm the errors using the information available about the events. These analyses for runoff QAQC included only 3 sub-watersheds, which will be expanded to all sub-watersheds in the future. So far, these approaches brought to light 813 rainfall and 24 runoff events with erroneous timestamps or magnitude of observations etc.

Acknowledgments

We thank the group of USDA-ARS dedicated technicians that for more than 60 years have professionally

maintained the network of rain gauges, runoff structures, and general experimental watershed infrastructure used in this study and numerous research projects. Rainfall, runoff, sediment, meteorology and watershed characterization data are freely available from <https://www.tucson.ars.ag.gov/dap/>. This research was a contribution from the Long-Term Agroecosystem Research (LTAR) network.

Data Availability Statement: The data that support the findings of this study are publicly available <https://www.tucson.ars.ag.gov/dap/>.

References

- Amitai, E., C. L. Unkrich, D. C. Goodrich, E. Habib, and B. Thill. 2012. “Assessing satellite-based rainfall estimates in semiarid watersheds using the USDA-ARS Walnut Gulch gauge network and TRMM PR.” *J. Hydrometeorol.* 13 (5): 1579–1588. <https://doi.org/10.1175/JHM-D-12-016.1>.
- Bitew, M. M., et al., 2019. Multi-parameter regression modeling for improving the quality of measured rainfall and runoff data in densely instrumented watersheds. *Journal Hydrologic Engineering*.
- Brakensiek, D.L., H. B. Osborn, and W. J. Rawls coordinators, 1979. *Field Manual for Research in Agricultural Hydrology*. U. S. Department of Agriculture, Agriculture Handbook No. 224, 550 pp., illus. Washington D.C.
- Brantley, S. L., et al., 2017. Designing a network of critical zone observatories to explore the living skin of the terrestrial Earth. *Earth Surface Dynamics*. 5, 841–860.
- Collinge, S. K., 2018. NEON is your observatory. *Frontiers in Ecology and the Environment*. 16, 371–371.
- Costa, A. C., A. Bronstert, and J. C. de Araujo. 2012. “A channel transmission losses model for different dryland rivers.” *Hydrol. Earth Syst. Sci.* 16 (4): 1111–1135. <https://doi.org/10.5194/hess-16-1111-2012>.
- Das, N. N., B. P. Mohanty, M. H. Cosh, and T. J. Jackson. 2008. “Modeling and assimilation of root zone soil moisture using remote sensing observations in Walnut Gulch Watershed during SMEX04.” *Remote Sens. Environ.* 112 (2): 415–429. <https://doi.org/10.1016/j.rse.2006.10.027>.
- Duan, Y., P. Heilman, and D. P. Guertin. 2006. “A web-based tool for economic analysis of sediment control on rangeland watersheds.” In *Proc., IEMSS 3rd Biennial Meeting: “Summit on Environmental Modelling and Software”*. Burlington, VT: International Environmental Modelling and Software Society
- Fiebrich, C. A., Crawford, K. C., 2001. The impact of unique meteorological phenomena detected by the Oklahoma Mesonet and ARS Micronet on automated quality control. *Bulletin of the American Meteorological Society*. 82, 2173–2187.
- Garcia, M., et al., 2008. Spatial interpolation of precipitation in a dense gauge network for monsoon storm events in the southwestern United States. *Water Resources Research*. 44.
- Goodrich, D.C., Lane, L.J., Shillito, R.M., Miller, S.N., Syed, K.H., Woolhiser, D.A. 1997. Linearity of basin response as a function of scale in a semiarid watershed. *Water Resour. Res.* 33(12):2951–2965.
- Goodrich, D. C., et al., 2008. Long-term precipitation database, Walnut Gulch Experimental Watershed, Arizona, United States. *Water Resources Research*. 44.
- Goodrich, D. C., I. S. Burns, C. L. Unkrich, D. Semmens, D. P. Guertin, M. Hernandez, S. Yatheendradas, J. R. Kennedy, and L. Levick. 2012. “KINEROS2/AGWA: Model use, calibration, and validation.” *Trans. ASABE* 55 (4): 1561–1574. <https://doi.org/10.13031/2013.42264>.
- Goodrich, D.C., D. G. Williams, C. L. Unkrich, J. F. Hogan, R. L. Scott, K. R. Hultine, D. Pool, A. L. Coes, and S. N. Miller. 2004. “Comparison of methods to estimate ephemeral channel recharge, Walnut Gulch, San Pedro River Basin, Arizona.” In *Recharge and Vadose Zone processes: Alluvial basins of the Southwestern United States*. Water Science and Application No. 9, edited by F. M. Phillips, J. F. Hogan, and B. Scanlon, 77–99. Washington, DC: American Geophysical Union.

- Keefer, T. O., et al., 2008. An event-based comparison of two types of automated-recording, weighing bucket rain gauges. *Water Resources Research*. 44.
- Keefer, T.O., Renard, K.G., Goodrich, D.C., Heilman, P., Unkrich, C.L. 2015. Quantifying Extreme Precipitation Events and their Hydrologic Response in Southeastern Arizona. *Journal of Hydrologic Engineering*. 21(1): 1-10. 10.1061/(ASCE)HE.1943-5584.0001270
- Kleinman, P. J. A., Spiegel, S., Rigby, J. R., Goslee, S. C., Baker, J. M., Bestelmeyer, B. T., ... & Duncan, E. W. (2018). Advancing the sustainability of US agriculture through long-term research. *Journal of environmental quality*, 47(6), 1412-1425.
- Knapp, A. K., et al., 2012. Past, Present, and Future Roles of Long-Term Experiments in the LTER Network. *Bioscience*. 62, 377-389.
- Knipper, K. R., T. S. Hogue, K. J. Franz, and R. L. Scott. 2017. "Downscaling SMAP and SMOS soil moisture with moderate-resolution imaging spectroradiometer visible and infrared products over southern Arizona." *J. Appl. Remote Sens.* 11 (2): 026021. <https://doi.org/10.1117/1.JRS.11.026021>.
- Kolassa, J., et al. 2018. "Estimating surface soil moisture from SMAP observations using a neural network technique." *Remote Sens. Environ.* 204 (Jan): 43–59. <https://doi.org/10.1016/j.rse.2017.10.045>.
- Kunkel, K. E., et al., 2005. Quality control of pre-1948 cooperative observer network data. *Journal of Atmospheric and Oceanic Technology*. 22, 1691-1705.
- Moran, M.S., Emmerich, W.E., Goodrich, D.C., Heilman, P., Holifield Collins, C., Keefer, T.O., Nearing, M.A., Nichols, M.H., Renard, K.G., Scott, R.L., Smith, J.R., Stone, J.J., Unkrich, C.L., Wong, J.K. 2008a. Preface to special section on Fifty Years of Research and Data Collection: U.S. Department of Agriculture Walnut Gulch Experimental Watershed. *Water Resources Research*, Vol. 44, W05S01.
- Moran, M. S., C. D. Holifield Collins, D. C. Goodrich, J. Qi, D. T. Shannon, and A. Olsson. 2008b. "Long-term remote sensing database, Walnut Gulch Experimental Watershed, Arizona, United States." *Water Resour. Res.* 44 (5): W05S10. <https://doi.org/10.1029/2006WR005689>.
- Morin, E., Krajewski, W.F., Goodrich, D.C., Gao, X., Sorooshian, S. 2003. Estimating rainfall intensities from weather radar data: The scale-dependency problem. *J. Hydrometeorology* 4:782-797.
- Niu, G., P. Troch, C. Paniconi, R. L. Scott, M. Durcik, X. Zeng, T. Huxman, D. C. Goodrich, and J. Pelletier. 2014. "An integrated modelling framework of catchment-scale ecohydrological processes. 2: The role of water subsidy by overland flow on vegetation dynamics in a semi-arid catchment." *Ecohydrology* 7 (2): 815–827. <https://doi.org/10.1002/eco.1405>.
- Nichols, M.H., Renard, K.G., Osborn, H.B. 2002. Precipitation changes from 1956-1996 on the Walnut Gulch Experimental Watershed. *J. Am. Water Resources Assoc.* 38(1):161-172.
- Osborn, H. B., L. J. Lane, and V. A. Myers. 1980. "Rainfall/watershed relationships for southwestern thunderstorms." *Trans. ASAE* 23 (1): 82–87. <https://doi.org/10.13031/2013.34529>.
- Renard, K.G., Lane, L.J., Simanton, J.R., Emmerich, W.E., Stone, J.J., Weltz, M.A., Goodrich, D.C., Yakowitz, D.S. 1993. Agricultural impacts in an arid environment: Walnut Gulch studies. *Am. Institute of Hydrology, Hydrological Sci. and Tech.* 9(1-4):149-159.
- Sansom, J., 1992. Breakpoint representation of rainfall. *Journal of Applied. Meteorology.*, **31**, 1514–1519.
- Scott, R. L., and J. A. Biederman. 2017. "Partitioning evapotranspiration using long-term carbon dioxide and water vapor fluxes: New approach to ET partitioning." *Geophys. Res. Lett.* 44 (13): 6833–6840. <https://doi.org/10.1002/2017GL074324>.
- Smith, R. E., D. L. Chery, K. G. Renard, and W. R. Gwinn. 1982. Supercritical flow flumes for measuring sediment-laden flow. *USDA ARS Technical Bulletins* 1655. Washington, DC: US Dept. of Agriculture,

Agricultural Research Service.

Spiegel, S., et al., 2018. Evaluating strategies for sustainable intensification of US agriculture through the Long-Term Agroecosystem Research network. *Environmental Research Letters*. 13, 034031.

Stone, J. J., M. H. Nichols, D. C. Goodrich, and J. Buono. 2008. "Longterm runoff database, Walnut Gulch Experimental Watershed, Arizona, United States." *Water Resour. Res.* 44 (5): W05S05. <https://doi.org/10.1029/2006WR005733>.

Syed, K., Goodrich, D.C., Myers, D., Sorooshian, S. 2002. Spatial characteristics of thunderstorm rainfall fields and their relation to runoff. *J. Hydrology* 271(1-4):1-21.

Tetzlaff, D., et al., 2017. The essential value of long-term experimental data for hydrology and water management. *Water Resources Research*. 53, 2598-2604.

Walbridge, M. R., Shafer, S. R., A Long-Term Agro-Ecosystem Research (LTAR) Network for Agriculture. The Fourth Interagency Conference on Research in the Watersheds, Alaska, 2011.

You, J., et al., 2007. Performance of quality assurance procedures on daily precipitation. *Journal of Atmospheric and Oceanic Technology*. 24, 821-834.

Tables

Table 1. Pros and cons of the four methods implemented for the runoff QA-QC

Figures

Figure 1 . a) Geographical location of WGEW. Average total summer rainfall for the period 2000-2017 and spatial distribution of the 99 rain gauges operational in the watershed. The thicker lines show the drainage boundaries for sub-watersheds WS-04, WS-06, and WS-011. Finer lines show the boundaries for other larger sub-watersheds b) USDA digital weighting rain gauge, c) sub-watershed WS-06 flume.

Figure 2 . Flow chart used to flag inconsistencies in daily, interpolated rainfall maps. Type I Errors are errors due to turned off rain gauges or missing gauge while Type II Errors are errors resulting from incorrect date coding.

Figure 3 . Schematic description of the four approaches used for runoff QAQC. a) Rainfall-runoff association method, b) Lag time, c) Runoff coefficient, d) Regression method, and e) plots of runoff and corresponding rainfall.

Figure 4: The example above shows questionable events identified by lag time method and visual aid to identify the value of threshold lag time for WS-11. The questionable events are identified as the events having outlier lag time values. 70 min is considered as a threshold lag time for this particular watershed where it showed clear separation from the rest of computed values.

Figure 5. Rainfall errors identified with the aid of rainfall maps, a) Type I and b) Type II.

Figure 6. Cumulative relative frequency of days with errors in the rainfall dataset from 1954 to 2014 (a and b) and the percentage contribution of each source of error to total error (c and d). The horizontal line in a and b indicates the date when the analog system was replaced by the digital system.

Figure 7. a) Percentage of events flagged by the four methods for the three sub-watersheds, b) percentage of the total number of flagged events identified by 1 to 4 methods simultaneously.

Figure 8. Selected runoff (black line) and rainfall from multiple gauges (gray lines) within the watershed of interest on inverted y-axis. The solid blue line represents conditional mean of the different rain gauges in the watersheds. The figure shows excessively delayed runoff in a and c, runoff events not associated with the corresponding rainfall in d and b, and runoff occurring before the rainfall was being recorded in f and e.

Hosted file

HYP_DAP_Meles_Table.pdf available at <https://authorea.com/users/362857/articles/483847-curating-62-years-of-walnut-gulch-experimental-watershed-dataset-improving-qaqc-methods-of-rainfall-and-runoff-measurements>

Hosted file

HYP_DAP_Meles_Figure.pdf available at <https://authorea.com/users/362857/articles/483847-curating-62-years-of-walnut-gulch-experimental-watershed-dataset-improving-qaqc-methods-of-rainfall-and-runoff-measurements>