Characterization of Hillslope Hydrologic Events Using Machine Learning Algorithms

Eunhyung Lee¹ and Sanghyun Kim¹

¹Department of Environmental Engineering, College of Engineering, Pusan National University, Busan, South Korea

Correspondence to: Sanghyun Kim (kimsangh@pusan.ac.kr)

Key Points:
- We analyzed rainfall and soil wetness dataset with machine learning algorithms.
- Seven delineated clusters of hydrologic events represent distinct hydrological processes.
- Representative soil moisture monitoring points can be obtained using the machine learning methods.
Abstract

Time series of soil moisture were measured at 30 points for 396 rainfall events on a steep, forested hillslope between 2007 and 2016. We then analyzed the dataset using an unsupervised machine learning algorithm to cluster the hydrologic events based on the dissimilarity distances between weighting components of a self-organizing map (SOM). Generation patterns of two primary hillslope hydrological processes, namely, vertical flow and lateral flow, at the upslope and downslope areas were responsible for the distinction of the hydrologic events. Two-dimensional spatial weighting patterns in the SOM provided explanations for the relationships between rainfall characteristics and hydrological processes at different locations and depths. High reliability in hydrologic classification was achieved for both the driest and wettest events; as assessed through k-fold cross validation using 10 years of data. Representative soil moisture monitoring points were found through temporal stability analysis of the event structure delineated from the machine learning classification. Application of a supervised machine learning algorithm provided a scheme using soil moisture for the cluster identification of hydrologic event even without rainfall data which is useful to configure hillslope hydrologic process with the least cost in data acquisition.

Keywords: rainfall, soil moisture, hillslope hydrology, temporal stability, machine learning
1 Introduction

Information on soil moisture is critical for assessing water storage, estimating the amount of runoff generated, and determining slope stability for hillslopes during rainfall (Angermann et al., 2017; Lu and Godt, 2008; Penna et al., 2011; Tromp van Meerveld and McDonnel, 2005). Hillslope hydrological processes are affected by many factors including topographic, soil textural, and eco-hydrologic parameters (Baroni et al., 2013; Liang et al., 2011; Rodriguez-Iturbe et al., 2006; Rosenbaum et al., 2012; Wilson et al., 2004), which results in highly nonstationary and heterogeneous temporal and spatial distributions of soil moisture (Penna et al., 2009; Wilson et al., 2004). The relationship between precipitation and runoff is highly nonlinear, and the spatio-temporal variations in soil moisture, groundwater, and surface runoff are extremely difficult to predict (Ali et al., 2013; Curti et al., 2014).

Temporal stability has been widely used for selecting representative points for the characterization of soil moisture variation (Minet et al., 2013; Vachaud et al., 1985). Temporal stability depends on soil depth, soil properties, land use, surface & subsurface topography and hydrometeorological conditions (Gao and Shao, 2012; Gao et al., 2015). Both the mean and standard deviation of the relative difference had been used to evaluate the temporal stability of soil moisture (Zhao et al., 2010). High stability is an important criterion for determining the best location for the monitoring spatially averaged soil moisture of a given area (Brocca et al., 2010; Penna et al., 2013; Ran et al., 2017). The location of spatially representative soil moisture points can be also explored in the context of process-based interpretations (Lee and Kim, 2017).

Rainfall is the primary driver of rapid variations in soil moisture and subsurface flow generation (Penna et al., 2011). The response of soil moisture to rainfall events has been investigated for different topographic positions, depth profiles, and land cover conditions (Feng and Liu, 2015; He et al., 2012; Wang et al., 2013; Zhu et al., 2014). The functional relationship between rainfall events and soil moisture varies and depends on various factors such as soil texture, depth, topography, and vegetation cover (Liang et al., 2011; Bachmair et al., 2012; Kim, 2016). Various rainfall characteristics including the total amount, duration, intensity, and dry period duration also have been used to understand the soil moisture response (Alberson and Kiely, 2001; Heisler-White et al., 2008). Other studies of rainfall features have categorized rainfall events for the analysis of soil moisture variation (Lai et al., 2016; Wang et al., 2008).

The generation of distinct hillslope flow paths; vertical flows such as matrix flow and bypass flow and lateral flows along different boundaries (e.g., subsurface stormflow over bedrock and surface overland flow) can be
differently appeared between upslope and downslope. The functional relationship between rainfall and soil water storage had been studied (Brocca et al., 2005; Castillo et al., 2003; Xie and Yang, 2013), but how the rainfall features such as rainfall amount, intensity, duration and antecedent soil moisture condition influence hydrological processes and their distributions at the hillslope scale had not been explored yet. Studies in hillslope hydrology had focused several events to identify specific flow path (e.g., subsurface lateral flow) using intensively collected field measurements for relatively short periods (Freer et al., 2004; Kim 2009; Penna et al., 2011; Wienhöfer and Zehe, 2013). A comprehensive approach can be explored to address the holistic behavior of hydrological processes using an extended dataset. Pattern recognition capabilities of dataset have proved useful for evaluating the dissimilarity of hydrologic measurements between different hydrologic events. The self-organizing map (SOM) method has been used to investigate datasets representing ecosystems, animals, catchment classification, and crop evapotranspiration (Adeloye et al., 2011; Farsadnia et al., 2014; Ley et al., 2011; Liu et al., 2011; Park et al., 2003). The SOM can be an effective method to understand a big hydrologic data through reducing dimensionality of dataset which can provide hydrologic interpretation. The highly heterogeneous and extremely nonstationary variation in soil moisture between the upslope and downslope areas as well as the upper and lower soil layers of a hillslope can be analyzed by using an SOM.

Machine learning techniques have been applied to soil moisture data from both in-situ measurements (Van Arkel and Kaleita, 2014) and remote sensing applications (Ahmad et al., 2009; Prashant et al., 2013). Supervised learning algorithms were used to improve predictions of subsurface flow in a hillslope (Bachmair and Weiler, 2012), to downscale satellite soil moisture data (Prashant et al., 2013), and to estimate soil moisture obtained through regression analysis (Ahmad et al., 2009). Identification of critical soil moisture sampling points has also been done using an unsupervised learning algorithm (Van Arkel and Kaleita, 2014). Most studies involving the use of machine learning algorithms for the analysis of soil moisture have focused on estimating and determining the appropriate measurement locations for assessing the variations in mean soil moisture.

In this study, we aimed to answer the following research questions:

1. How can machine learning algorithms be used to understand the soil moisture response patterns at the hillslope scale?

2. Can delineated clusters of hydrologic events be explained by different hillslope hydrological processes?
3. How can the representative soil moisture monitoring points be obtained through supervised machine learning to determine the variation of mean soil moisture, and to distinguish hydrologic events?

In this study, an alternative understanding of hillslope hydrologic behavior is explored through a long term data analysis with machine learning. Hydrologic events for the hillslope scale can be characterized through rigorous classification of a big hydrologic dataset. In particular, machine learning algorithms provide several opportunities for understanding hydrologic events through the transformation of a substantial dataset into compact clusters and for delineating the hierarchical relationship between clusters, which can be useful for exploring process-based interpretations and for obtaining an efficient monitoring network. We used a hydrologic data (throughfall and soil moisture) to analyze and characterize the highly complex relationships between antecedent soil moisture, rainfall characteristics, and soil moisture responses. An unsupervised neural network method, namely, a self-organizing map (SOM), was introduced to investigate the nonlinear interactions between various rainfall characteristics and their effects on temporal change in soil moisture, and to classify the multivariate datasets in terms of the likely flow paths in the hillslope. A supervised machine learning algorithm, namely, a C4.5 decision tree (Quinlan, 1993), was then employed to obtain optimal soil moisture points to characterize rainfall events; thus, we were able to efficiently identify the hydrologic events with less soil moisture sensor data and determine the mean soil moisture variation.

To address these research topics, we employed the following research approaches. First, we applied an SOM algorithm to datasets composed of throughfall and soil moisture distributions from upslope to downslope locations in the study area. The dataset was reclassified based on the weighting vectors of each neuron in the SOM map by using the Euclidean distances between 10 hydrologic variables from different hydrologic events.

Secondly, the nonlinear relationship between throughfall and soil moisture was evaluated by comparing the spatially weighted patterns of rainfall characteristics and soil wetness variables. The relationships between particular throughfall characteristics and soil moisture at different depths and locations were investigated, and these data were used for interpretations of hydrological processes.

Thirdly, representative sampling points of soil moisture for each distinct hydrologic classification were identified. The decision scheme made with selected soil moisture points showed the potential for using a supervised machine learning algorithm to identify hydrological processes with minimal monitoring costs.
2 Materials and Methods

2.1 Study Area and Data Acquisition

The study hillslope (4000 m²) is located in the Sulmachun watershed (8.5 km²), which is a headwater of the Imjin River in northwestern South Korea (Figure 1). The study area is primarily covered by a mixture of Polemoniales, shrubby Quercus, and a coniferous canopy of Pinus densiflora, and the slope varies between 30° and 45°. Rainfall, streamflow, and other hydrometeorological records (e.g., temperature and relative humidity) have been collected over the last 25 years from seven hydrologic monitoring stations in this watershed (see Figure 1). Mean annual rainfall for the last two decades was approximately 1,500 mm; 70% of the total rainfall fell during the Asian monsoon season between June and August. Precipitation occurred as snowfall between December and March. Mean annual evaporation was approximately 420 mm, estimated with the eddy-covariance method by using data obtained from a flux tower (adjacent hydrologic monitoring station) located 50 m away from the study area; monthly evaporation exceeded the accumulated rainfall only in October 2010. Average daily temperature varied between −15 and 35°C. The hillslope bedrock consists of granite with extensively weathered areas. Elevations range between 200 and 260 m above sea level, and the surface slope varied between 20° and 35°. Leptosol and Cambisol (classifications from the Food and Agricultural Organization of the United Nations (FAO)) are the dominant soils of the upslope and downslope areas, respectively. Analysis of 15 soil samples (five points each for the upslope, middle slope, and downslope areas at depths of 30 cm) indicated that the predominant soil textures are sandy loam and loamy sand. The average porosities were 49% and 48% for the upslope and downslope areas, respectively. Multiple insertions of an iron pole to each grid cell (0.5 m by 0.5 m) indicated that the soil depth along the hillslope varies between 25 and 95 cm. The depth of the root zone is approximately 20 to 30 cm.

Throughfall (this was used to describe rainfall characteristics) was recorded at hourly intervals by using a rainfall gauge (Automatic Rain Gauge System, Eijkelkamp) located under the canopy. The soil moisture time series were measured using a multiplex-based time domain reflectometer (TDR; MiniTRASE, SoilMoisture, 2004) at 5 locations upslope (UP1–UP5) and 5 locations downslope (DO1–DO5) (Figure 1). At each location, three TDR sensors (waveguides) were inserted parallel to the surface at depths of 10, 30, and 60 cm into the upslope side of an installation trench. Soil moisture measurements were collected bi-hourly between 2007 and 2016. There were 396 rainfall events...
during the study period. The rainfall event was defined by the minimum duration of a dry period as 1 day and the minimum amount of throughfall as 1 mm.

2.2 Temporal Stability

The temporal stability method can be applied to soil moisture datasets to evaluate the temporal variability in soil moisture (Vachaud et al., 1985). The normalized difference between soil moisture and mean soil moisture can be expressed as

$$\delta_{i,j}(\%) = \frac{\theta_{i,j} - \overline{\theta}_j}{\overline{\theta}_j} \cdot 100$$

where $\theta_{i,j}$ is the measured soil moisture, at location $i$ and time $j$, respectively, and $\overline{\theta}_j$ is the spatially averaged soil moisture for time $j$, which can be estimated as $\overline{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} \theta_{i,j}$, where $n$ is the number of soil moisture measurements for the hillslope.

The index of temporal stability (ITS$_i$) was proposed to quantify the soil moisture stability at point $i$ (Zhao et al., 2010). A high ITS$_i$ is indicative of low temporal stability, whereas a low ITS$_i$ is indicative of high temporal stability. The value of ITS$_i$ can be calculated as follows:

$$ITS_i(\%) = \sqrt{MRD_i^2 + SDRD_i^2} \cdot 100$$

The mean relative difference (MRD$_i$) can be calculated as $MRD_i = \frac{1}{m} \sum_{j=1}^{m} \delta_{i,j}$, where $m$ denotes the final measurement time. The standard deviation of the relative difference (SDRD$_i$) represents the variability in the normalized difference for the monitoring period, which can be calculated as follows:

$$SDRD_i(\%) = \sqrt{\frac{\sum_{j=1}^{m} (\delta_{i,j} - MRD_i)^2}{m-1}} \cdot 100$$

2.3 Unsupervised Machine Learning Algorithm

The SOM is an unsupervised learning algorithm that can be useful for pattern recognition for a multivariate dataset from different observations. The SOM is typically a two-dimensional (2D) grid composed of either hexagonal...
or rectangular elements. In this study, we used a hexagonal lattice as the output layer because it resulted in better feature of information propagation when updating more neighborhood neurons than that of the rectangular lattice (Kohonen, 2001). Based on the recommended output dimension of $5\sqrt{r}$ (Kohonen, 2001), where $r$ is the number of events, and the 396 total rainfall events used in this study, the array structure of the SOM was specified as a $16 \times 6$ matrix, which corresponded to 96 neurons, i.e., the grid cells in the SOM. Each neuron had a different weighting vector ($w_{ab}$), where the subscripts $a$ and $b$ represent address codes for the variable and node, respectively. A random number was used to initialize the weighting vectors in neurons.

For a given rainfall event, the soil moisture variation at a particular point in the hillslope depends not only on the rainfall, but also on other environmental factors such as the location, depth, and soil texture. In order to consider the relative variation (%) of water storage normalized by the antecedent moisture condition, we used the percentage of maximum soil moisture difference (Zhu et al., 2014) as an index, namely, the soil moisture difference index, to represent soil moisture variation:

$$\Delta \theta(\%) = \frac{\theta_{\text{max}} - \theta_{\text{ant}}}{\theta_{\text{ant}}} \cdot 100$$  

where $\theta_{\text{max}}$ is the maximum soil moisture during a rainfall event and the subsequent period ($\leq 4$ h), and $\theta_{\text{ant}}$ is the soil moisture measurement before the rainfall event (2 hours).

Once the dataset is populated with the throughfall characteristics and soil moisture data, the SOM map (Kohonen, 2001) can be obtained. In this study, throughfall duration, total throughfall amount, and mean throughfall intensity were used as the rainfall characteristics, and the antecedent moisture content, soil moisture difference for 5 locations upslope at depths of 10, 30, and 60 cm, and soil moisture difference for 5 locations downslope at depths of 10, 30, and 60 cm were used for the soil moisture data (Figure 1). Because each variable had a different range, we applied a natural logarithm transformation out of several transformations such as Box-Cox transformations with different parameters to the dataset standardize the variables, which centralize the means of all variables into zero.

The SOM maps were established to each variables and the distance between the input vector and weighting vector can be calculated as follows:

$$d_b = \sqrt{\sum_{a=1}^{v} (w_{a,b} - x_a)^2}$$  

where $v$ is the number of variables.
The winner neuron can be identified as the neuron with the minimum value of $d_b$ indicating the best fitness to characteristics of each rainfall event among every neurons in SOM map. If the winner neuron is chosen, the weighting vector should be re-evaluated by using the Eq. for the renewal weighting vector as follows:

$$\Delta w_{a,b} = \begin{cases} \alpha(x_a - w_{a,b}) & b = b^* \\ 0 & b \neq b^* \end{cases}$$

$$w_{a,b}^{new} = w_{a,b}^{old} + \Delta w_{a,b}$$

where $\alpha$ ($= 0.5$) is an acceleration coefficient and $b^*$ is the winner neuron. Neurons adjacent to the winner neuron are also updated through the application of Eq. (6). Both the radius of neighboring neurons and acceleration coefficient decreased (from 16 to 1) linearly as the iterations approached the maximum iteration.

After finishing the work with the updating algorithm, all neurons in the SOM maps had fitted weighting vectors to the multiple datasets of this study. The probability density function of each input variable leading to selection of each specific SOM node can then be inferred from the weighting vector. The input variables in each neuron can be displayed in the form of a spatial pattern in the SOM maps. They have many kinds of spatial patterns, which correspond to the number of variables. Additionally, the displays for one variable and many variables are called the “component plane” and “component planes,” respectively. The nonlinear relationship between variable was identified through visual comparison between the spatially distributed weightings in each component plane (Adeloye et al., 2011; Farsadnia et al., 2014; López García and Machón González, 2004; Park et al., 2003).

2.4 Clustering

Clusters within the dataset can be delineated by applying the dendrogram classification method and by evaluating the dissimilarity between the weighting vectors (Montero and Vilar, 2014). In this study, we used a hierarchical method because the resulting dendrogram structure provided better representation for the relationships between clusters than the results obtained by using non-hierarchical methods. The hierarchical method forms clusters by binding datasets with shorter distances between them. The Euclidean distance was employed to evaluate the dissimilarity because it is suitable for shape-based comparisons between soil moisture series collected at the same time (Iglesias and Kastner, 2013). This method also used to identify clusters of soil moisture data (Van Arkel and
The Euclidean distance between two weighting vectors in the neurons ($b_1$ and $b_2$) can be expressed as follows:

$$d_{b1b2} = \left[ \sum_{a=1}^{v} (w_{a,b1} - w_{a,b2})^2 \right]^{0.5}$$ (7)

The relationship that has the shortest distance between neurons is assigned to the first cluster, and weighting vectors of the first cluster can be expressed as

$$\mu_{c1,a} = \frac{n_{b1}\mu_{b1} + n_{b2}\mu_{b2}}{n_{b1} + n_{b2}}$$ (8)

where $\mu_{b1}$ and $\mu_{b2}$ are the variable weighting vectors in the neurons ($b_1$ and $b_2$), respectively; $n_{b1}$ and $n_{b2}$ are set to 1 in this relationship, but these values are set to the numbers of components during the comparisons of clusters. Additionally, we used Ward’s method to evaluate the dissimilarity between two weighting vectors of each neuron, and between each cluster, i.e., this was the chosen algorithm in our hierarchical clustering method (Ward, 1963). When the dissimilarity between two clusters ($c_1$ and $c_2$) is calculated, the distance between the clusters can be expressed as

$$d_{cluster} = \sum_{a=1}^{v} \frac{||\mu_{a,c1} - \mu_{a,c2}||^2}{n_{c1} + n_{c2}}$$ (9)

where $\mu_{a,c1}$ and $\mu_{a,c2}$ are the averages of clusters $c_1$ and $c_2$, respectively, and $n_{c1}$ and $n_{c2}$ are the numbers of components for clusters $c_1$ and $c_2$, respectively. A dendrogram can be constructed based on the resulting $d_{cluster}$, and the upper part from a designated horizontal line can be recognized as the structure of the final clusters.

2.5 Water Storage Evaluation

The water storage after a rainfall event can provide a better understanding about what flow paths or combined processes contribute to the redistribution of soil water along the hillslope. The water storage can be evaluated following a multiplication step for the corresponding soil depth (i.e., 200 mm for soil moisture at the depths of 10 and 30 cm, and 300 mm for soil moisture at the depth of 60 cm) related to the maximum difference of soil moisture between prior and post rainfall conditions. The distribution of water storage in depth profile and hillslope location (upslope or downslope) indicate distinct hydrological processes along the hillslope. The collected throughfall data can be used for the effective rainfall while neglecting evaporation for the short period of analysis.
One of the most reliable supervised learning algorithms is the C4.5 algorithm (Quinlan, 1993), which can be used for regression or classification of multivariate data. The C4.5 algorithm is based on the selection of the attribute that yields the maximum information gain from data subsets. The information gain can be calculated as follows:

\[ H(X) = \sum_{x \in X} p(x) \log_2 \left( \frac{1}{p(x)} \right) \]  

where \( H(X) \) is the entropy of all clusters, which is a measure of uncertainty, \( p(x) \) is the proportion of element numbers in each cluster, and \( X \) is the number of clusters. In other words, the entropy for cluster delineations from SOM can be obtained from equation (10).

In order to obtain the threshold of soil moisture difference index, the information gain is introduced. The higher information gain represents the better threshold for cluster identification. The information gain is a measure of the difference in prior and post entropy for the set, which is split on the attribute \( A \) as

\[ IG(A,X) = H(X) - \sum_{t \in T} p(t) H(t) \]  

where \( IG(A,X) \) is the information gain from the classification of attribute \( A \) (i.e., \( A \) is the criteria for the specific soil moisture difference at a designated point), \( T \) is either greater than \( A \) or less than \( A \), and \( t \) is a subset of \( T \). The term in equation (11), \( p(t) \), is composed of two conjugate probabilities; the events have smaller soil moisture difference index then the threshold divided by total events and the events have greater soil moisture difference index then the threshold divided by total events.

The C4.5 algorithm can be used to identify a specific soil moisture point that can be used to classify a designated cluster by calculating the entropy. Thresholds for all soil moisture monitoring point are evaluated and the one having highest information gain is selected as a classifier of decision tree and the identical procedure is repeated to obtain the C4.5 decision tree of soil moisture difference indices. Depending on the soil moisture characteristics for several points, the main feature of the hydrologic system can be identified. A decision tree, namely, a sequential structure of soil moisture criteria, can be developed for the specific locations of soil moisture with the maximum information gain. In other words, the soil moisture variation of the selected measurement points can be used to delineate identical clusters obtained by the dendrogram classification method.
3 Results

3.1 Representative Point for Soil Moisture Measurement

The ITS of soil moisture was estimated for 10 locations (Figure 1(b)) at three soil depths (10, 30, and 60 cm) to determine a representative point for analyzing the soil moisture variation for rainfall events. The ITSs for soil moisture measurements at 30 points showed that both the location and depth affected the stability of soil moisture (Figure 2). The ITS tended to be higher at deeper depths because of the impact of hydrometeorological drivers (e.g., rainfall and evaporation) and intermittent generation of subsurface storm flow at deeper depths, which corresponds to the results from a previous study (Lee and Kim, 2017). Location UP5 had the lowest ITS of the 10 locations, and thus, it represents the most temporally stable soil moisture (at all depths). UP3-10 had the lowest ITS of all 30 points; specifically, the value was 0.13. Figure 3 shows the soil moisture at both the most temporally stable point (UP3-10) and the least temporally stable point (DO5-30) for representing the average soil moisture of all points. As shown in Figure 3, RMSEs (root mean square errors) were 2.22% for UP3-10 and 8.72% for DO5-30. The representative point, which had the lowest ITS, was correlated more with the average soil moisture than any other point. Therefore, the soil moisture measurement for UP3-10 was adopted as the representative soil moisture before the event for the SOM analysis.

3.2 Composition and Clustering of the SOM

With the application of the SOM method, the dataset of hydrologic measurements (396 × 10) was transformed through 96 neurons and output in terms of a matrix (16 × 6) through the iterative application of Eqs. (5) and (6). In other words, 10 hydrologic variables from 396 events were expressed compactly in the SOM.

Dissimilarity in terms of the Euclidean distance between output neurons was then used to construct the dendrogram. Many alternatives exist to number of clusters depending on the complexity of the dendrogram structure. In this study, seven clusters was selected based on a heuristic approach aiming to achieve a hydrologically meaningful classification of events and parsimonious clustering. The relation to notable hydrological processes such as lateral flow or vertical preferential flow and the redundancy check in cluster number were important factors in the heuristic approach. Figure 4(a) shows the resulting dendrogram for the seven clusters. The structure of the dendrogram shows
the relationships between groups of clusters and between individual clusters. For example, the relationship between clusters 5 and 6 had a lower hierarchy than that of clusters 3 and 4. Figure 4(b) presents the output SOM (16 × 6) delineated from the dendrogram analysis, which is a structural array identical to the delineated dendrogram with neurons for each cluster. The spatial distributions between other clusters and corresponding numbers of neurons indicate the areal portion of each cluster from all clusters and its connection with adjacent clusters. Table 1 presents the average vector components (e.g., duration of rainfall (DUR), amount of throughfall (AMO), intensity of throughfall (INT), antecedent soil moisture (ASM) in volumetric %, UP10, UP30, UP60; average of the soil moisture difference indices (Δθ) in five upslope locations at a depth of 10 cm, 30 cm, and 60 cm, respectively, DO10, DO30, and DO60; average of the soil moisture difference indices (Δθ) in five downslope locations at depths of 10 cm, 30 cm, and 60 cm, respectively) for the seven clusters shown in Figures 4(a) and 4(b).

As shown in Figure 4(b), clusters 1 and 2 were located in the upper part of the SOM. Table 1 indicates that the rainfall characteristics of clusters 1 and 2, such as DUR, AMO, and INT, were relatively low and that the antecedent soil moisture was similar to the mean ASM for all clusters (Figure 5). The average soil moisture difference indices were less than 5% for cluster 1 because the low throughfall amount and intensity resulted in a limited increase in soil water storage, and the loss due to evaporation offset a substantial proportion of the precipitation (Albertson and Kiely, 2001; Ramirez et al., 2007). Cluster 1 was most similar to cluster 2 but had higher throughfall amounts and intensities (see Figure 4(a)). The intermediate part of the SOM (Figure 4(b)) included clusters 3 and 4, which had higher throughfall durations, amounts, and intensities than clusters 1 and 2. The soil moisture difference indices for clusters 3 and 4 were higher than those for clusters 1 and 2. The higher throughfall durations and amounts for cluster 4 were associated with higher soil moisture difference indices for cluster 4, but the lower throughfall for cluster 3 resulted in smaller soil moisture difference indices (Table 1). One notable feature of clusters 3 and 4 was the increasing trend of soil moisture difference indices with depth (DO60 > DO30) for the downslope area, whereas those of clusters 1 and 2 showed decreased soil moisture difference indices with depth (DO30 > DO60) (Table 1).

The events showing similarity with larger soil moisture difference index, namely, significant events, in the SOM were in clusters 5, 6, and 7, as shown in Figure 4(b). The pattern of soil moisture difference indices for cluster 5 was indicative of vertical infiltration upslope and strong lateral flow downslope (Table 1 and Figure 4), which were distinct from cluster 6. The soil moisture difference indices for cluster 5 were larger than those of cluster 6, except for UP60 (9.6%), even with the lower throughfall duration, amount, intensity, and antecedent soil moisture than those for
cluster 6. Both rainfall characteristics and soil moisture difference indices for cluster 7 were significantly higher than those for all other clusters. Many of the measurement points in cluster 7 were saturated throughout rainfall events, and the DO10 value of cluster 7 was 192.6% (48.2% in terms of volumetric soil moisture), which indicates that overland flow occurred was likely at the downslope points.

3.3 Water Storage Analysis

We selected a rainfall event from each of the seven clusters that showed the distinct characteristics of hydrological processes associated with each cluster for the analysis. The water storage distribution analysis for seven rainfall events is presented in Table 2. Water storage analysis for clusters 1 and 2 showed that negligible changes in water storage occurred for both the upslope and downslope areas, and the main difference between cluster 1 and cluster 2 was whether the rainfall affected the soil moisture difference index (%) at a depth of 60 cm in the upslope area (Table 2). Rainfall impacts to clusters 3 and 4 were classified into the intermediate category because both clusters introduced meaningful storage change (mm) and soil moisture difference index (%) in the deepest depths of the downslope area (60 cm) indicative of the generation of subsurface lateral flow. Significant changes in water storage (>40 mm) were found for clusters 5, 6, and 7 regardless of the amount of throughfall. The main difference between clusters 5 and 6 was whether the vertical preferential flow affected the 60 cm soil layer in the upslope area, which depended on the antecedent soil moisture. Cluster 7 was mainly associated with very high amounts of rainfall, and subsurface lateral flow was generated in both the upslope and downslope areas; additionally, exceptionally large changes in water storage (>116 mm) occurred in the downslope area.

3.4 Component Planes of Each Variable

The component planes of 10 variables and their visual comparisons can provide a better understanding of the nonlinear relationships between the 10 hydrological variables. Figures 5(a)–5(j) show the vector component weightings of the 10 variables. Both the spatial distributions and the scales of weightings (scale bar) in Figure 5 represent the characteristics of impacts (rainfall characteristics and antecedent soil moisture) and consequences (soil moisture difference). As shown in Figures 5(a), 5(f) and 5(g), higher weightings for the throughfall amount,
throughfall duration, and throughfall intensity were located in the lower-right part of the SOM map. The SOM map for antecedent soil moisture showed dry conditions in the lower-left part and wet conditions in the lower-central portion (Figure 5(b)). Visual comparison between rainfall features and antecedent soil moisture revealed that the correlation between throughfall and antecedent soil moisture was negligible. Therefore, the periods between the rainfall events were long enough on average that the influence of previous rainfall events on soil moisture variation was negligible.

Figure 5(c) shows the weightings of soil moisture difference indices upslope at a depth of 10 cm, which is the soil layer influenced by direct precipitation with negligible impacts from the upslope area. The combined throughfall amount weightings (Figure 5(a)) and throughfall intensity weightings (Figure 5(f)) appeared similar to the weightings in Figure 5(c). Similar weighting distributions were observed for the upslope area at depths of 30 cm and 60 cm (Figures 5(d) and 5(e)) with degreasing trend in deeper depth, which corresponds to cluster 5. Weightings of soil moisture difference indices for the downslope area (Figures 5(h)–5(j)) showed less variation in the horizontal direction in the lower part than those for the upslope area indicating the behaviors of clusters 5, 6, and 7 are similar in downslope. This means that the soil moisture variations for the downslope area were less correlated to rainfall characteristics than those for the upslope area. The soil moisture difference index map for the downslope area at a depth of 10 cm (Figure 5(h)) was correlated only to the throughfall intensity (Figure 5(f)). The soil moisture difference indices at depths of 30 cm and 60 cm (Figures 5(i) and 5(j)) showed distributions that were less horizontally skewed in terms of the weighting than those for the depth of 10 cm (Figure 5(h)). The impact of rainfall and antecedent soil moisture appeared to be dampened at greater depths downslope. No notable correlation was found between ASM (Figure 5(b)) and soil moisture difference indices.

3.5 Validation of the SOM Classification

The reliability of the SOM application for hydrologic events was also evaluated through comparison of the SOM made from the complete dataset with those from partial datasets (k-cross validation, k = 10 years). Multiple SOMs were made with datasets missing events in 2016, 2015, ..., and 2007, which resulted in 10 SOMs from incomplete datasets for a 10-year period. SOMs from partial datasets had identical dendrogram structures to that of the complete dataset, which resulted into seven clusters shown in Figure 4. The numbers of neurons for each of the
seven clusters cannot be identical between the 10 SOMs. To evaluate the robustness of the SOM characterization, different missing events for each of the 10 incomplete datasets were projected to each of the corresponding 10 SOMs made from partial datasets. In other words, unused datasets for each of the 10 training datasets were used to reclassify neurons and cluster numbers by evaluating the Euclidean distances between the weighting vectors. Figure 6 presents the matrix expression between the event projections for the SOM shown in Figure 4(b) and the sum of event projections from the 10 SOMs made from partial datasets. For clusters 1, 5, and 7 in the matrix (Figure 6), there was more than 90% agreement in event numbers between the complete SOM and the other partial classifications, whereas cluster 2 resulted in 49% agreement in terms of the classification, which was mainly related to misidentifications between cluster 2 and cluster 1. The number of agreements from k-cross validation was 317 out of 396 hydrologic events in 10 years.

3.6 Supervised Machine Learning for Soil Moisture Representative Points

The stability of soil moisture has been widely assessed to obtain representative points for studying the variation in mean soil moisture (Brocca et al., 2010; Ran et al., 2017; Vachaud et al., 1985). Unlike in the ITS analysis, the soil moisture monitoring points for effective discrimination between different clusters should be sensitive to rainfall events, and also effectively address hydrological processes such as infiltration and soil water redistribution along the hillside. A decision-making process for cluster identification can be expressed in terms of a decision tree. Thus, the supervised machine learning algorithm was applied to construct a C4.5 decision tree (Quinlan, 1993).

As shown in Figure 7, the C4.5 decision tree consisted of soil measurement points in the upslope (UP3-10, UP5-10, UP4-60, UP5-60) and downslope (DO3-10, DO4-10, DO5-10, DO2-60) areas. From the delineated decision tree (Figure 7), the point at DO2-60 provided the greatest information gain from all points because soil moisture difference index at this point determined the distinctions between clusters 6 and 7, clusters 4 and 5, clusters 1 and 2, and clusters 4 and 5. A point at UP5-60 contributed to the discrimination between cluster 3, cluster 4, and cluster 5 from cluster 6 and cluster 7, and a point at DO5-10 was used to distinguish between cluster 2 and cluster 3. Furthermore, point UP4-60 was important for determining the difference between cluster 1 and cluster 2, while point DO3-10 was important for determining the difference between cluster 3 and cluster 4.
4 Discussion

4.1 Hydrological Processes

Application of an unsupervised machine learning algorithm, namely, SOM, to the dataset provided an integrated assessment to evaluate and characterize hydrologic events. Then, soil moisture recharge could be inferred based on the rainfall characteristics and antecedent moisture content. In other words, the recharge patterns of water storage for soil layers of the hillslope were characterized by several distinct clusters. The hydrologic events were classified into three distinct categories, which depend on the generation subsurface lateral flow in downslope (discussing in the following section), and seven further refined clusters as follows: insignificant events for clusters 1 and 2, intermediate events for clusters 3 and 4, and significant events for clusters 5, 6, and 7. Further classification of significant events indicated that the effects of antecedent moisture conditions and throughfall amounts were critical to delineating cluster 5 and cluster 7. The generation of lateral flow in deep soil layers of the downslope area was considered to be the threshold feature between the insignificant and intermediate events; the primary difference between the intermediate events and significant events was the substantial development of interface flow between bedrock and soil layer in the downslope area (see Figure 8(c)).

As illustrated in Table 2, the distinct distribution of soil water storage can be explained by the different combinations and degrees of hydrological processes (vertical flow, subsurface lateral flow, and preferential vertical flow) for each cluster. The comparison between cluster 5 and clusters 4 and 6 indicated that there was strong nonlinearity in the generation of hydrological processes (stronger lateral subsurface flow in the downslope area for the events of cluster 5 even with small throughfall and less antecedent soil moisture than those for clusters 4 and 6). This means that even though throughfall and antecedent soil moisture are important in determining hydrological processes for each cluster, the generation of different hydrological processes cannot be completely explained by these factors alone and the nonlinearity of hydrological responses needs to be explored.

The hillslope hydrological flow path was characterized through the comparison of component planes between UP10 and UP30 or UP60, and other combinations of soil moisture component planes, such as those of DO10 and DO30 or DO60. The exclusive vertical flow impact could be identified based on relationships between the component plane for UP10 and those for UP30 or UP60 (Figures 5(c), 5(d), 5(e)) because there was small upslope contributing areas or topographic wetness indices (Figure 1) in upslope locations (Beven and Kirkby, 1979). The high weightings
of 10 cm for the upslope area were distributed in two parts of SOM, namely, the lower-left and lower-right (Figure 5(c)), but those of deeper depths for the upslope area were found only on the lower-right part of SOM (see Figures 5(d) and 5(e)). Weightings in UP10 were associated with the throughfall amount and throughfall intensity, but those for UP30 and UP60 were correlated only to the throughfall amount. A consistent pattern of weighting was found between UP30 and UP60 compared to UP10, which was attributed to the effect of vertical infiltration. Furthermore, the higher weightings tended to decrease as depth increased in the upslope area because the effect of vertical infiltration was smaller at greater depths (Li et al., 2013). The relationships between the component plane of DO10 and those of DO30 and DO60 differed (Figures 5(h)–5(j)). Even though high weightings in the middle-left part for DO10 partially decreased in the component planes for DO30 and DO60, high weightings in the lower-left corner in Figure 5(h) remained in Figures 5(i) and 5(j). The spatial weighting patterns of soil moisture at downslope points (Figures 5(i) and 5(j)) did not show any notable correlations to those of upslope soil moisture at identical depth and rainfall characteristics. This means that the flowpath in the downslope area cannot be completely explained by vertical flow. Furthermore, the component plane in Figure 5(h) different from that of Figure 5(c), which means that the soil moisture at a depth of 10 cm in the downslope area could have been also affected by the upslope contributing area.

The weighting ranges (scale bars in Figure 5) of the 10 and 60 cm depths (Figures 5(h) and 5(j)) were greater than that of the 30 cm depth (Figure 5(i)). This may indicate that the lateral flow along boundaries (subsurface and surface) was stronger than that at intermediate depths (Table 1). Cluster 6 and cluster 5 were located in the lower part of the SOM map, as shown in Figure 4(b). As illustrated in Table 1, cluster 6 had higher amounts, durations, and intensities of throughfall than cluster 5, and antecedent soil moisture for clusters 6 and 5 were the highest and lowest, respectively, among all clusters. The soil moisture difference indices for cluster 6 were similar or greater than 50% at all locations except DO30 (32.1%) (Table 1), which indicates that downslope lateral flow tended to be generated through boundaries either along the surfaces or bedrock (Kim, 2009). Furthermore, rainfall and antecedent soil moisture were substantially higher for cluster 6 than for cluster 5, and the soil moisture difference indices in cluster 6 were relatively more uniform for all points than those for cluster 5 (Table 1). This may be explained by the development of preferential pipe flow, which is more common at greater depths in wetter conditions (Lai et al., 2016; Uchida et al., 2001; Wienhöfer and Zehe, 2013). Low variation in UP60 for cluster 5 indicated that low antecedent moisture conditions limited active lateral flow into the downslope area. Extreme events were mainly associated with cluster 7, as illustrated in Table 1. Lateral flow likely occurred in both in the upslope and in the downslope areas for
cluster 7. In particular, Figures 8(b) and 8(c) show statistics for volumetric soil moisture for cluster 7 that were indicative of the substantial development of saturation. As shown in Figures 5(e)–5(j), the soil moisture changes both UP30 and DO30 were lower than UP10 or UP60 and DO10 or DO60, respectively. Effective drainage during extreme events seemed to be strongly associated with lateral flow generation along the two boundaries in the soil media (i.e., surface and bedrock) (Freer et al., 2004; Haga et al., 2001; Kim, 2009). The impact of extreme rainfall conditions dominated over other controls (e.g., land cover and topography) in terms of hillslope runoff generation (Feng and Liu, 2015).

4.2 Reliability in Hydrologic Event Classification

The contribution of each vector component to the SOM classification shown in Figure 4(b) was analyzed through the distribution of statistical characteristics for each cluster, as shown in Figures 8(a)–8(c). As shown in Figure 8(a), the impact of ASM on the soil moisture classification was distinctive for clusters 6, 4, and 7. Even though cluster 5 and cluster 7 were adjacent to each other in the dendrogram structure, the ASM values for the two clusters were completely different, which indicates that ASM could not have been the dominant factor for soil moisture incremental changes during extreme events. Box plots of cluster 5 in relation to the throughfall duration and throughfall amount were also substantially different from those of adjacent clusters such as cluster 6 and cluster 7 (see Figure 8(a)). Box plots of volumetric soil moisture (VSM) by cluster generally showed increasing distributions from cluster 1 to cluster 7 in terms of both the mean values and variance, except for cluster 5, both in the upslope and downslope areas (Figures 8(b) and 8(c)). In order to test the reliability of the SOM classification, we used the k-cross validation technique for 10 different datasets in 10 years. Depending on the characteristics of rainfall events each year, the degree of agreement of the SOM projections between the complete dataset and the partial datasets differed. A comprehensive evaluation of the SOM predictability was performed through comparison between the event projections of the complete dataset SOM and the summation of the projections of the 10 partial SOMs (Figure 6). All disagreement was due to missed recognition between adjacent clusters (Figure 4(b)) in the SOMs from partial datasets, as shown in Figure 6. Even though the cluster identification of partial dataset SOMs between adjacent clusters did not always perfectly match that of the complete dataset, the predictability of extreme events was high and stable. Actually, the degree of agreement
for three categories was 92.9%. This demonstrates the reliability of the SOM for the characterization of hydrologic events of big data.

4.3 Temporal Stability and Representative Point

Considering the importance of determining the representative point for soil moisture monitoring, the classification of hydrologic events can be approached in terms of the temporal stability of soil moisture (Minet et al., 2013; Penna et al., 2013; Vachaud et al., 1985). A representative point can be also designated for all clusters through ITS analysis (Table 3). The selected points shown in Table 3 were the most temporally stable points for each cluster. The representative point for clusters 1, 2, and 5 was UP3-10 for all events, but clusters 3, 4, 6, and 7 yielded a different point (DO3-10) as the representative point. The differences in mean soil moistures and distinct hydrological processes (generation of vertical flow and lateral flow in the upslope area) throughout hydrologic events between clusters 1, 2, 5 and clusters 3, 4, 6, 7 seemed to be responsible for these two different representative points. The statistics were also different between the two representative points. The $R^2$ and RMSE values were evaluated between the averages of the soil moisture time series and those of a representative point in each cluster. As illustrated in Table 3, DO3-10 provided a higher $R^2$ than UP3-10 for clusters 3 and 4, but the RMSEs for UP3-10 were lower than those for UP3-10 for clusters 3 and 4, and the opposite was found for cluster 7. The $R^2$ values were identical between UP3-10 and DO3-10, DO3-10 provided lower RMSEs than UP3-10 for cluster 6. The results shown in Table 3 indicate that the two points (UP3-10 and DO3-10) can be used as representative points for the seven clusters.

The representative point for soil moisture monitoring can differ depending on the soil, depth, topography, and vegetation (Bachmair et al., 2012; Baroni et al., 2013; Gao et al., 2015; Zhu et al., 2014). Furthermore, temporal stability can be sensitive to the temporal distribution of rainfall (Penna et al., 2013). We classified 396 hydrologic events into seven clusters, and the temporal stability analysis for each cluster shown in Figure 4(a) resulted in only two points for stable variations in soil moisture. This effectiveness of representative points was partially related to the fact that the soil moisture stability analysis was performed under similar hydrologic conditions based on the identified hydrological processes shown in Table 3. The spatial and profile distributions of vertical flows and lateral flows can be comprehensively characterized through 7 clusters noted in Table 3. In other words, the soil moisture monitoring in...
2 representative points (one in upslope and the other in downslope) provides a holistic configuration of flow paths as well as the evaluation of mean soil moisture variation in the study area.

4.4 Representative Points for Cluster Identification

On the basis of further analysis in regard to the hydrological processes shown in Table 3, a compact set of soil moisture monitoring points was selected and used to identify the seven clusters (Figure 7). As shown in Figure 7, whether soil moisture difference index at DO2-60 was less than 9.2 or not was the first criterion prior to moving to the next step such as determining whether the soil moisture difference index at UP5-10 was less than 5.2 or that of UP5-60 was less than 16.4. One or two more steps further lead to the identification of the clusters (Figure 7).

Unlike in the ITS analysis, the monitoring point for the effective distinction between different clusters should be sensitive to rainfall events. Including the two representative points (UP3-10 and DO3-10) for the upslope and downslope areas selected from the ITS analysis, eight points were used in the C4.5 decision tree (Figure 7). The soil moisture points in the C4.5 decision tree did not include any measurement point at a depth of 30 cm. This indicated that the important hydrological process (lateral flow) for event distinction was mainly generated at either surface or bedrock boundaries.

The hydrological processes indicated in Table 3 can be expressed in terms of a decision tree (Figure 9). A substantial similarity was found between Figures 7 and 9 in terms of the orders of clusters between the two decision trees. One or multiple diverging branches of the soil moisture difference index decision tree (Figure 7) corresponded to one of the diverging branches of the decision tree for hydrological processes (Figure 9). This was because the soil moisture response, whether it was greater than or less than the threshold soil moisture difference index (Figure 7), was the combined result from multiple hydrological processes acting at the corresponding point.

The comparison between Figures 7 and 9 indicated that soil moisture difference index for DO2-60 could be used to determine whether hydrologic events were significant or “other” (insignificant or intermediate) events, as well as the degree of downslope hydrological processes for clusters 5, 6, and 7. The soil moisture difference index at point UP5-60 was useful in identifying generations of vertical flow and lateral flow in the upslope area. The point DO5-10 contributed to determinations of the degree of vertical flow in the upslope area and differences in the rainfall categories.
between insignificant and significant events. The existence of vertical flow in the upslope area was detected in the soil moisture difference index at UP4-60.

Even though the decision tree in Figure 7 provided an optimum monitoring set for cluster identification, the accuracy of cluster identification was less than 100%. This was due in part to our use of partial data (26%) to predict the behavior of the total dataset. The other possible explanation seemed to be related to the underlying stationary assumption of the hydrologic system for the study period (10 years). The difference in canopy activity over 10 years and the generation of catastrophic rainfall events (e.g., 145 mm in 2 hours on 27 July 2011) could have partially changed the redistribution mechanisms of soil moisture along the study area. The accuracy of the delineated decision tree (Figure 7) in identifying the seven clusters was 82% and that of the rainfall category classification (insignificant, intermediate and significant) was 95%. However, eight soil moisture monitoring points even without rainfall and antecedent soil moisture presented in Figure 7 demonstrated a reliable capacity for making distinctions among clusters constructed from the total dataset (rainfall characteristics, antecedent moisture, and 30 points of soil moisture difference index for 396 events) and the representative points of mean soil moisture variation in both the upslope and downslope areas. In other words, the capability of dimensionality reduction of the machine learning algorithm is useful not only in the data grouping of similar behavior with hydrologic interpretation but also in the delineating of the minimum monitoring points for the cluster identification, which can be useful to substantially reduce the cost of network maintenance.

5 Conclusions

Rainfall characteristics and the responses of soil moisture at the hillslope scale were explored through the application of machine learning algorithm to a big dataset of hydrologic responses. Hydrologic events were characterized through the application of an unsupervised learning algorithm to a soil moisture dataset collected over 10 years from a steep hillside. Based on a delineated dendrogram, classification of neurons into seven clusters and three primary event types provided meaningful interpretations to understand the hydrologic events. The upslope and downslope spatial patterns of hillslope hydrological processes, vertical flow, and lateral flow were responsible for the distinctions between the event clusters. The nonlinear relations between hydrologic variables were expressed effectively in 2D SOM presentations of variables. Comprehensive tests of the SOMs with 10 partial datasets and
projections of missing datasets revealed the robustness of the application of the SOM for the classification and
prediction of extreme hydrologic events. Water storage analysis for each event from the seven clusters suggest that
different combinations and contributions of vertical flow, subsurface lateral flow, and preferential flow determined
the particular hydrological process dominant for each cluster. Temporal stability analysis of soil moisture time series
provided efficient representative points for all delineated clusters. The soil moisture decision tree obtained from the
application of a supervised learning algorithm effectively identified the clusters of hydrologic events even without
abundant rainfall data and antecedent soil moisture data. The application of learning machine algorithms was useful
not only to understand soil moisture variation patterns within clustered events, but also to identify the optimal
monitoring locations for the mean soil moisture variation and different generations of vertical flow and lateral flow in
upslope and downslope. The approach developed in this study should be applicable to other hydrological systems
having sufficient data with connectivity in processes between variables.

Code and Data Availability

Code and data will be available through repository https://www.re3data.org/ when paper is accepted.

Author contribution
Enhyung Lee and Sanghyun Kim and several former graduate students had collected data for the study area. Eunhyung
Lee developed model code and performed simulation. Sanghyun Kim prepared manuscript with contribution from
Enhyung Lee.

Competing interests
The authors declare that they have no conflict of interest

Acknowledgments
This study was financially supported by the Basic Research Program (2016R1D1A1B02008137) of the Korea
Research Foundation of the Republic of Korea.

References


Figure 1. Location of the Sulmachun watershed in South Korea with hydrologic monitoring (rainfall and streamflow) stations (lower left) and study area with terrain contours, the topographic wetness index (TWI) (Beven and Kirkby, 1979), and soil moisture monitoring points (right).
Figure 2. Indices of temporal stability (ITS) for 10 locations at depths of 10, 30, and 60 cm.
**Figure 3.** Soil moisture measurements for representative points (UP3-10) and the least temporally stable point (DO5-30) for the average soil moisture from 30 measurement points.
Figure 4. Structure of the (a) dendrogram for seven clusters and (b) SOM classifications in 96 neurons through a $16 \times 6$ matrix.
Figures 5. (a)–(j) Component planes of variable weightings for the rainfall amount, antecedent soil moisture (ASM), soil moisture difference indices for the upslope at depths of 10, 30, and 60
cm, rainfall duration, rainfall intensity, and soil moisture difference indices for the downslope at depths of 10, 30, and 60 cm.

**Figure 6.** 2D array expression of event projections for the summation of 10 SOMs of partial datasets to the SOM of the complete dataset.
Figure 7. A C4.5 decision tree of soil moisture difference indices (numbers in boxes) for hydrologic event classification.
Figure 8. Statistical distributions of component vectors of (a) rainfall characteristics upslope, (b) volumetric soil moisture, and (c) downslope volumetric soil moisture for the SOM classification.
Figure 9. A decision tree of hydrological processes shown in Table 3; partial means that the corresponding process was generated partially in the designated area; full indicates that the hydrological process was generated at all depths and locations.
Table 1. Arithmetic averages of vector components for the rainfall duration (DUR), throughfall amount (AMO), throughfall intensity (INT), antecedent soil moisture (ASM), means of the soil moisture difference index at depths of 10, 30, and 60 cm for upslope points (UP10, UP30, and UP60, respectively), and means of the soil moisture difference index at depths of 10, 30, and 60 cm for downslope points (DO10, DO30, and DO60, respectively).

<table>
<thead>
<tr>
<th>Cluster</th>
<th># of Event</th>
<th>DUR h</th>
<th>AMO mm</th>
<th>INT mm/h</th>
<th>ASM %</th>
<th>UP10 %</th>
<th>UP30 %</th>
<th>UP60 %</th>
<th>DO10 %</th>
<th>DO30 %</th>
<th>DO60 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>127</td>
<td>6.3</td>
<td>2.1</td>
<td>0.5</td>
<td>16.0</td>
<td>3.8</td>
<td>2.2</td>
<td>2.2</td>
<td>4.5</td>
<td>3.6</td>
<td>2.0</td>
</tr>
<tr>
<td>2</td>
<td>69</td>
<td>8.1</td>
<td>5.0</td>
<td>0.9</td>
<td>15.8</td>
<td>10.7</td>
<td>4.8</td>
<td>6.3</td>
<td>16.3</td>
<td>7.3</td>
<td>4.6</td>
</tr>
<tr>
<td>3</td>
<td>51</td>
<td>13.8</td>
<td>11.7</td>
<td>1.5</td>
<td>16.3</td>
<td>31.7</td>
<td>8.8</td>
<td>8.0</td>
<td>29.3</td>
<td>13.1</td>
<td>14.0</td>
</tr>
<tr>
<td>4</td>
<td>33</td>
<td>16.9</td>
<td>19.0</td>
<td>1.6</td>
<td>16.9</td>
<td>45.4</td>
<td>19.4</td>
<td>11.1</td>
<td>57.1</td>
<td>23.2</td>
<td>29.3</td>
</tr>
<tr>
<td>5</td>
<td>27</td>
<td>12.4</td>
<td>19.5</td>
<td>1.8</td>
<td>9.8</td>
<td>102.0</td>
<td>42.9</td>
<td>9.6</td>
<td>124.0</td>
<td>78.1</td>
<td>132.2</td>
</tr>
<tr>
<td>6</td>
<td>25</td>
<td>20.4</td>
<td>38.7</td>
<td>2.3</td>
<td>19.5</td>
<td>55.7</td>
<td>49.9</td>
<td>46.9</td>
<td>56.5</td>
<td>32.1</td>
<td>57.7</td>
</tr>
<tr>
<td>7</td>
<td>64</td>
<td>29.3</td>
<td>85.2</td>
<td>3.4</td>
<td>16.5</td>
<td>107.0</td>
<td>83.0</td>
<td>99.7</td>
<td>192.6</td>
<td>110.8</td>
<td>129.3</td>
</tr>
</tbody>
</table>
Table 2. Water storage analysis of selected rainfall events for all clusters.

<table>
<thead>
<tr>
<th>Year/Month</th>
<th>Throughfall Impact</th>
<th>Cluster</th>
<th>Antecedent Soil Moisture</th>
<th>Throughfall (mm)</th>
<th>storage change (mm)</th>
<th>soil moisture difference index (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>07/10</td>
<td>negligible</td>
<td>1</td>
<td>19.80</td>
<td>5.17</td>
<td>2.60</td>
<td>UP10 4.50</td>
</tr>
<tr>
<td>13/06</td>
<td></td>
<td>2</td>
<td>13.50</td>
<td>6.50</td>
<td>3.20</td>
<td>UP30 0.62</td>
</tr>
<tr>
<td>10/09</td>
<td></td>
<td>3</td>
<td>23.10</td>
<td>13.40</td>
<td>4.00</td>
<td>UP60 1.20</td>
</tr>
<tr>
<td>16/07</td>
<td></td>
<td>4</td>
<td>18.70</td>
<td>22.70</td>
<td>7.60</td>
<td>DO10 3.60</td>
</tr>
<tr>
<td>14/10</td>
<td></td>
<td>5</td>
<td>10.50</td>
<td>20.00</td>
<td>1.00</td>
<td>DO30 0.80</td>
</tr>
<tr>
<td>11/07</td>
<td></td>
<td>6</td>
<td>20.50</td>
<td>35.00</td>
<td>2.20</td>
<td>DO60 1.50</td>
</tr>
<tr>
<td>11/06</td>
<td></td>
<td>7</td>
<td>8.60</td>
<td>85.50</td>
<td>14.60</td>
<td></td>
</tr>
</tbody>
</table>

Impact: negligible, intermediate, significant.
Table 3. Representative points of all clusters with coefficients of determination ($R^2$) and root mean square errors (RMSEs), and hydrological processes for all clusters.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Representative Point</td>
<td>UP3-10</td>
<td>UP3-10</td>
<td>DO3-10</td>
<td>DO3-10</td>
<td>UP3-10</td>
<td>DO3-10</td>
<td>DO3-10</td>
</tr>
<tr>
<td>Statistics</td>
<td>$R^2$</td>
<td>0.919</td>
<td>0.943</td>
<td>0.946</td>
<td>0.805</td>
<td>0.637</td>
<td>0.721</td>
</tr>
<tr>
<td>RMSE</td>
<td>1.547</td>
<td>1.705</td>
<td>1.851</td>
<td>1.880</td>
<td>2.110</td>
<td>1.568</td>
<td>3.132</td>
</tr>
<tr>
<td>rainfall impact</td>
<td>insignificant</td>
<td>Intermediate</td>
<td>Significant</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vertical</td>
<td>up</td>
<td>no</td>
<td>partial</td>
<td>full</td>
<td>full</td>
<td>partial</td>
<td>full</td>
</tr>
<tr>
<td>slope</td>
<td>no</td>
<td>surface</td>
<td>all</td>
<td>all</td>
<td>middle</td>
<td>all</td>
<td>all</td>
</tr>
<tr>
<td>flow</td>
<td>down</td>
<td>no</td>
<td>partial</td>
<td>full</td>
<td>full</td>
<td>full</td>
<td>full</td>
</tr>
<tr>
<td>slope</td>
<td>no</td>
<td>surface</td>
<td>all</td>
<td>all</td>
<td>all</td>
<td>all</td>
<td>all</td>
</tr>
<tr>
<td>lateral</td>
<td>up</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>No</td>
<td>no</td>
<td>surface</td>
</tr>
<tr>
<td>slope</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>No</td>
<td>no</td>
<td>bedrock</td>
<td>bedrock</td>
</tr>
<tr>
<td>flow</td>
<td>up</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>No</td>
<td>no</td>
<td>bedrock</td>
</tr>
<tr>
<td>down</td>
<td>partial</td>
<td>partial</td>
<td>partial</td>
<td>partial</td>
<td>partial</td>
<td>partial</td>
<td>partial</td>
</tr>
<tr>
<td>slope</td>
<td>no</td>
<td>no</td>
<td>bedrock</td>
<td>bedrock</td>
<td>bedrock</td>
<td>bedrock</td>
<td>bedrock</td>
</tr>
</tbody>
</table>

Partial: the corresponding flowpath was generated partially in the designated part; full: the process was generated at all depths and locations; surface: the process was generated at a depth of 10 cm; middle: the process was generated up to a depth of 30 cm; all: the process was generated at all depths; bedrock: the process was generated at a depth of 60 cm.