An Automated Anomaly Detection System for Hourly Rainfall Data Quality Control

Hway-Min Chou, Ke-Sheng Cheng

Master Program in Statistics, National Taiwan University

Abstract

Data quality assurance has been receiving increasing attention in the field of hydrology in the last decade. Only high-quality data ensures data-driven risk analysis and decision-making strategies of hydrology applications. In Taiwan, the Central Weather Bureau manages an automated rain gauge network system of over 600 stations to obtain real-time precipitation observations. Occasionally, rainfall observations of one station are markedly higher or lower than those of nearby stations, suggesting the presence of anomalies because rainfall observations of neighboring stations are often highly correlated. To obtain reliable results based on hourly rainfall data, these anomalies should be identified in advance. However, there is a lack of definite criteria for effectively identifying anomalies.

In this study, we established an automated anomaly detection system for precipitation observations. First, we categorized the data into four groups according to the four fundamental storm types in Taiwan. Second, we adopted K-means clustering analysis to classify all rain gauge stations of interest by their geographical location and rainfall characteristics. For each cluster, PCA was conducted to acquire the first few principal components, aiming to construct an index representing the extent of anomalies. Once the criteria are determined, identifying anomalies is straightforward. Eventually, we established the detection system and presented it as an online interactive web page. Therefore, a dependable anomaly detection system was created for effectively screening out possible anomalies to achieve hourly rainfall data quality control.

Keywords: Hourly Precipitation, Data Quality Control, Anomaly Detection, PCA, K-Means Clustering Analysis

1. Introduction

Rainfall data are essential to agricultural farming, travel planning, and performing nearly all daily activities. The Central Weather Bureau (CWB) manages an automated rain gauge network system of over 600 stations to obtain real-time precipitation observations in Taiwan. Countless decisions required for livelihood activities rely on the analyses of these rainfall observations. Accordingly, the quality of rainfall data is paramount, necessitating rainfall data quality assurance (QA) and rainfall data quality control (QC). Data QA investigates inconsistencies and anomalies in the original data. Data QC uses the information from the QA process to determine whether the data can be used for analysis or applications. QA approaches utilized in manufacturing have wide applications, including observation, data archiving, and processing and dissemination of environmental information (Hudson et al., 1999). In the field of hydrology, data QA has been received increasing attention (You et al., 2007; Branisavljević et al., 2009).

Occasionally, anomalies occur the hourly observations provided by rain gauge stations. For example, when a station fails to send the observations in time because of malfunctions, delays, or unknown reasons, the amount of delayed observation becomes exceptionally high because it has been accumulating for several hours. Moreover, the rainfall data returned by a station may be notably higher or lower than those reported by nearby stations, suggesting the presence of anomalies because the rainfall amounts of neighboring stations are often highly correlated. To guarantee the reliability of hourly rainfall data, these anomalies must be identified. However, no definite criteria exist for instantly and effectively discovering these anomalies, and manual identification would be inefficient and infeasible. Therefore, in this study, we established an automated anomaly detection system for hourly precipitation observations. Using this system, rainfall data QC can be accomplished in a cost-effective manner.

Toe et al. (2017) conducted K-means cluster analysis and principal component analysis (PCA) to investigate the spatial and temporal variation patterns in the Central Dry Zone (CDZ) of Myanmar. They considered the influence of the climatological monsoon break on precipitation in the CDZ. Additionally, they divided the stations into different clusters to reveal the orographic effect and distinct climate dynamics. Their data revealed that the first and second principal components (PCs) mainly accounted for the spatial variabilities and seasonal (temporal) variation in average monthly precipitation in the CDZ, respectively. Before employing PCA, Toe et al. (2017) performed clustering to classify the original stations. Stations belonging to the same cluster possess similar rainfall characteristics. Furthermore, the obtained PCs could fully capture both spatial and temporal variations in precipitation.

In this study, we used statistical methods to generate the criteria for identifying anomalies. Because rainfall amounts are greatly affected by various rainfall characteristics (Boyle & Chen, 1987; Chen et al., 1999; Chen & Chen, 2003), we grouped the stations of interest to identify anomalies. Inspired by the method of Toe et al. (2017), we conducted adopt K-means cluster analysis (Cox, 1957; Fisher, 1958) of the stations based on the features related to geographical locations and primary storm types in Taiwan (Wang & Cheng, 1982). Then, we performed PCA (Pearson, 1901; Hotelling, 1933; Jolliffe, 2002) for detecting outliers.

The rest of the paper is structured as follows. Section 2 presents data collection and preprocessing. Section 3 illustrates the methods used to identify nine categories of anomalies. The K-means clustering results and anomalies detected by PCA are presented and discussed in Section 4. Section 5 provides the conclusion.

2. Data

The hourly rainfall data recorded by 297 rain gauge stations set up by the CWB are used because they provide consistent rainfall data of better quality. The unit of each hourly rainfall is millimeter per hour. Next, we web-scraped the hourly rainfall data from January 1, 1998, to May 30, 2020, from the Central Weather Bureau Observation Data Inquire System (CODiS), which is an online open data platform that offers free observation data of CWB's automatic weather stations.

We then preprocessed the collected data according to the rainfall characteristics of Taiwan. Since rainfalls of Taiwan are readily affected by four main storm types of Taiwan: frontal rain, Meiyu, convective storms, and typhoons. The hourly rainfall data were divided into four groups according to the rainy seasons of these storm types, as they have different rainfall characteristics.

Table 1 presents the rainy seasons and duration of the four storm types. We easily separated frontal rain and Meiyu by their rainy seasons. However, both convective storms and typhoons tend to occur from July to October. To successfully distinguish the two events, we considered the duration of each rainfall event from July to October. Furthermore, we referred to the list of warning typhoons from 1998 to 2020 issued by the CWB. If the duration of a rainfall event exceeded 12 h and corresponded with a typhoon warning, hourly rainfalls of that event were classified as typhoons instead of convective storms.

Table 1.

Rainy Seasons and Duration for Four Storm Types

Storm Type	Rainy Season	Duration
Frontal Rain	Nov. - Apr.	> 1 h
Meiyu	May and June	-
Convective Storms	July - Oct.	1 - 12 h
Typhoons	July - Oct.	> 12 h

3. Methods

3-1. Abnormal Situations

When hourly rainfall data of a specific station are considerably lower or higher than those of neighboring stations, this may be an anomaly. Table **2** lists nine circumstances for a station in 1 day that may be abnormal and cause apparently different rainfall time series between a specific station and neighboring stations.

Table 2.

Codes Corresponding to Specific Circumstances for a Station in One Day

Code	Circumstance
B1	OBS at some hours were higher than that of nearby stations
B2	Observed trace; nearby stations, rainfalls
B3	Did not observe OBS due to malfunctions
B4	Did not observe OBS due to delays
B5	Observed rainfalls; nearby stations, trace
B6	Observed rainfalls; nearby stations did not due to malfunctions
B7	Observed rainfalls; nearby stations did not due to delays
B8	Delayed return of accumulated rainfall records
B9	Rainfall trend was different from that of nearby stations

Note.
Trace = an amount of precipitation that is ≤ 0.1 millimeter;
OBS = Observations.

3-2. K-Means Clustering Analysis

The anomaly is the marked difference between rainfall data of a station from those of nearby stations. To effectively detect this type of anomaly, we employed the K-means clustering method to classify 297 rain gauge stations because rainfall characteristics vary with diverse geographical location and storm type.

The K-means clustering method partitions a data set into K distinct and non-overlapping clusters. Before clustering, the desired K clusters need to be determined. Then, the algorithm allocates each observation to one of the K clusters. Assuming n observations in our data set, $C_1, C_2, \dots, C_K$ denotes sets that include the indices of observations in each cluster. The K-means clustering method aims to minimize the within-cluster variation among K clusters. The within-cluster variation of cluster C_k is denoted as $V(C_k)$, which yields the following equation:

$$\min_{C_1,\dots,C_k} \{\textstyle\sum_{k=1}^{K} V\,(C_k)\} \tag{1}$$

Then, $V(C_k)$ is defined using the squared Euclidean distance.

$$V(C_k) = \frac{1}{|C_k|} \textstyle\sum_{x_i \in C_k} \parallel x_i - \bar{x}_k \parallel^2 \tag{2}$$

where $|C_k|$ denotes the number of observations in the k^{th} cluster, and $\bar{x}_k$ is the mean of cluster C_k (also called the cluster centroid).

$$\min_{C_1,\dots,C_k} \{\textstyle\sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{x_i \in C_k} \parallel x_i - \bar{x}_k \parallel^2\} \tag{3}$$

The algorithm work to solves Equation (3):

I. Each observation is randomly allocated a number from 1 to K, which serves as the initial cluster assignment.
II. Iterations occur until the alteration of assignments stops:
    A. The centroid $\bar{x}_k$ is computed for each K cluster (i.e., the mean for the observations in cluster C_k).
    B. Each observation is allocated to the cluster whose centroid is the closest, as defined by the Euclidean distance.

3-3. Principal Component Analysis

Once the cluster analysis was finished, PCA was used to develop the criteria for the automatic system for detecting anomalies. PCA, a technique for summarizing the information of a data set, was developed by Pearson (1901), Hotelling (1933), and Jolliffe (2002). PCA reduces the dimensionality of multivariate data while preserving meaningful information as much as possible. It uses unsupervised learning, relying entirely on the input data itself instead of the corresponding target data. PCA transforms the original data to a new coordinate system. The new set of variables, known as PCs, is a linear transformation of the original variables. Each new variable is uncorrelated with other new variables. After projecting the initial data, the first coordinate lies in the direction with the largest variance, the second coordinate with the second largest variance, and so on. The equation of PCA is given by

$$\boldsymbol{Z} = \boldsymbol{\Phi X} \tag{4}$$

where $\boldsymbol{Z}$ denotes the PCs, $\boldsymbol{\Phi}$ is a matrix of coefficients called loads determined by PCA, and $\boldsymbol{X}$ is a data matrix with n observations and a set of p features. Equation (4) yields p linear transformations that form the PCs using the original variables. The first PC is written as

$$Z_1 = z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \ldots + \phi_{p1}x_{ip},\ i=1,\ 2, \ldots ,n \tag{5}$$

This has the largest sample variance ($Var(Z1)$ is maximum) and is subject to the constraint that $\sum_{j=1}^{p} \phi_{j1}^2 = 1$. Without the constraint, these elements can result in an arbitrarily large variance. The remaining Z_i values are computed such that their variances are maximized and subject to another constraint, so that the covariance between Z_i and Z_j ($i \neq j$) equals to 0. For example, the optimization problem is solved to obtain the first PC.

$$\max_{\phi_1, \ldots \phi_p} \sum_{i=1}^{n} z_{i1}^2 = \max_{\phi_1, \ldots \phi_p} \{\frac{1}{n} \sum_{i=1}^{1} (\sum_{j=1}^{p} \phi_{j1}\, x_{ij})^2\} \tag{6}$$

We calculated the matrix $\boldsymbol{\Phi}$ using the covariance matrix $\boldsymbol{S}$, which is written as follows:

$$s_{ij} = \frac{\sum_{k=1}^{n}(x_{ik} - \overline{x}_i)(x_{jk} - \overline{x}_j)}{n-1} \tag{7}$$

Therefore, the singular decomposition of $\boldsymbol{S}$ solves the PCA problem.

$$\boldsymbol{U}^T\boldsymbol{SU} = \boldsymbol{L} \tag{8}$$

where $\boldsymbol{L}$ is a diagonal matrix containing the eigenvalues of $\boldsymbol{S}$, and $\boldsymbol{U}$ is a matrix containing the eigenvectors of $\boldsymbol{S}$. $\boldsymbol{\Phi}$ can be computed by these two matrices.

$$\boldsymbol{\Phi} = \boldsymbol{UL}^{-\frac{1}{2}} \tag{9}$$

If we scale the variables and make their variances equal to one, then $\boldsymbol{\Phi}$ is simply the eigenvector matrix $\boldsymbol{U}$. The covariance matrix becomes a correlation matrix $\boldsymbol{R}$. When $\boldsymbol{S}$ is replaced with $\boldsymbol{R}$, the principal components can be calculated by

$$\boldsymbol{Z} = \boldsymbol{\Phi}^T\boldsymbol{D}^{\frac{-1}{2}}\boldsymbol{X} \tag{10}$$

where $\boldsymbol{D}$ is the diagonal matrix obtained by $\boldsymbol{S}$ with each s_{jj} equals to one.

3-3 Establishing the criteria for anomaly detection

We conduct PCA from the temporal variation aspect, aiming to find the temporal variation in rainfalls at each station. Given a specified cluster, the data matrix $\boldsymbol{X}$ of this cluster on one day is $\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1m} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nm} \end{bmatrix}$, where each column vector $X_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix}$ denotes the hourly rainfall of n raingauge stations at hour j (the length of j must be at least larger than two). Next, each variable X_j is normalized to obtain the correlation matrix $\boldsymbol{R}$. The original data set is normalized because PCA computes a novel projection based on the standard deviation of the variables. A variable with an extremely high standard deviation will be given a higher weight for composing the new axis than a variable with a low standard deviation. If we normalize the data set in advance, then every variable will retain the same weight. By using Equation (10), we gain the first and second PCs Z_1 and Z_2.

Thereafter, the Euclidean distance between the origin and X_j being projected on the PCA subspace of the first two PCs is calculated.

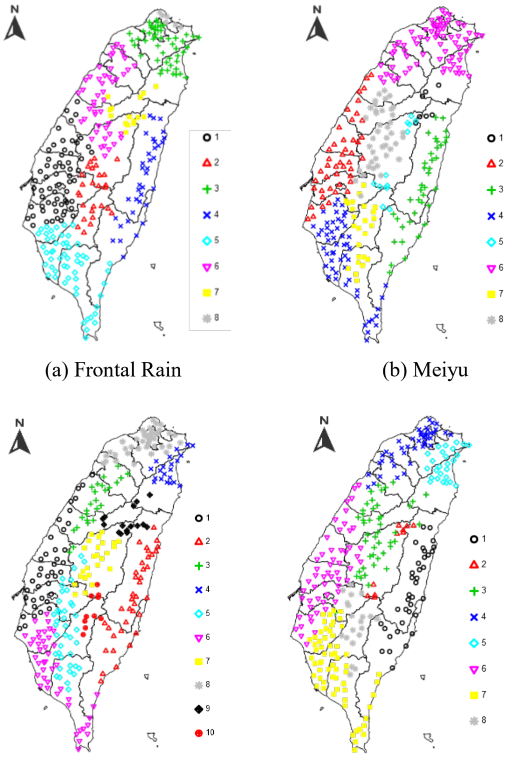
$$d = \sqrt{Z_1^2 + Z_2^2} \ \geq \ d_{i,j,p} \tag{11}$$

Considering n as the number of days in which PCA can be performed (it rained on these n days), the number of rainy days with i^{th} storm type and j^{th} cluster is $n_{i,j}$.

For these $n_{i,j}$ days, each day the maximum distance from the farthest projected data point to the origin can be computed. The $d_{i,j,p}$ is obtained by taking the p^{th} quantile of those maximum $n_{i,j}$ distances and set $d_{i,j,p}$ as the threshold for determining anomalies. If d (the Euclidean distance from any projected data point to origin) exceeds $d_{i,j,p}$, this suggests the existence of anomalies at a specific station because PC1 (Z_1) captures the largest spatial variation, and PC2 (Z_2) accounts for the remaining variation of those normalized variables. The temporal variation in rainfalls explained by each PC is non-overlapping.

4. Results

The ideal clustering results of four storm types are presented in Figure 1.



(a) Frontal Rain

(b) Meiyu

(c) Convective Storms

(d) Typhoons

Figure 1. Clustering Results for Four Storm Types

The following five variables for each storm type were calculated and used to conduct the K-means clustering analysis of 297 stations:

I. The altitude of a station.

II. The longitude of a station.

III. The latitude of a station.

IV. The average annual rainfall from 1998 to 2019 of a specified storm type.

V. The standard deviation of the average annual rainfall from 1998 to 2019 of a specified storm type.

Performing PCA from a temporal variation aspect enables us to observe the temporal variation patterns in rainfall for each rain gauge station. We took March 27, 2020, as an example. The data matrix of this day is a 36 × 9 matrix because Cluster 4 of the frontal rain type contains 36 stations (Table 3), and the 13th, 14th, 15th, 16th, 17th, 18th, 22th, 23th, and 24th h of this day observed rainfalls. After normalizing the data matrix and conducting PCA, we obtained the variable correlation plot (Figure 2) and the new coordinate system (Figure 3) formed by PC1 and PC2.

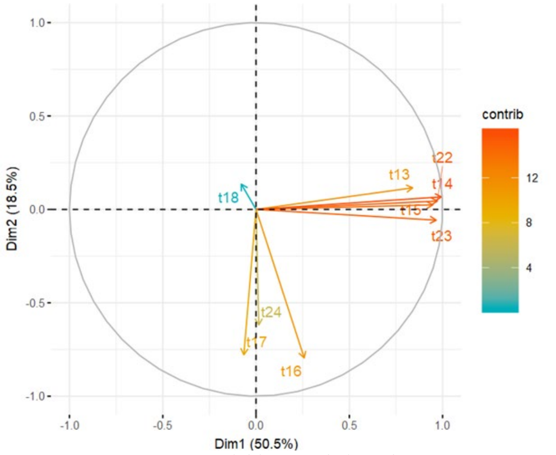


Figure 2. Variable Correlation Plot

For Figure 2, the horizontal axis represents PC1, which accounts for 50.5% variation in our original data matrix; the vertical axis represents PC2, accounting for 18.5% variation. Thus, the first two PCs explain 69% variation of the rainfall of this day. Figure 2 shows the correlation coefficients r between 2 PCs and the nine original variables, which can be obtained as

$$r = \frac{v_{ij} \times e_j}{Std(X_i)} \tag{12}$$

where v_{ij} denotes the i^{th} element of the j^{th} unit-length eigenvector of the covariance matrix, e_j denotes the eigenvalue of PC_j ($Var(PC_j)$), and $Std(X_i)$ denotes the standard deviation of the variable X_i. Because the data matrix is normalized, the value of $Std(X_i)$ is 1. Using Equation (12), the relationship between each PC and a specific variable can be obtained. For instance, the correlation coefficient of PC1 with the 13$^{\text{th}}$ hour is 0.84, whereas that of PC2 and the 13$^{\text{th}}$ hour is 0.12.

The colors in Figure 2 represent the expected contribution of a variable to the PCs. The contribution of a variable to a given PC (in percentage) is computed as follows:

$$contrib = \frac{r_{ij}^2 \times 100}{\sum_j \sum_i r_{ij}^2} \tag{13}$$

where r_{ij} denotes the correlation coefficient of variables X_i and PC_j. The expected contribution is attained using

$$\frac{\sum_j(contrib\times e_j)}{\sum_j e_j} \tag{14}$$

where e_j denotes the j^{th} eigenvalue (variance) of PC_j. For example, the contributions of the 13th hour to PC1 and PC2 are 15.51% and 0.86%, respectively, whereas the expected contribution is approximately 11.57%.

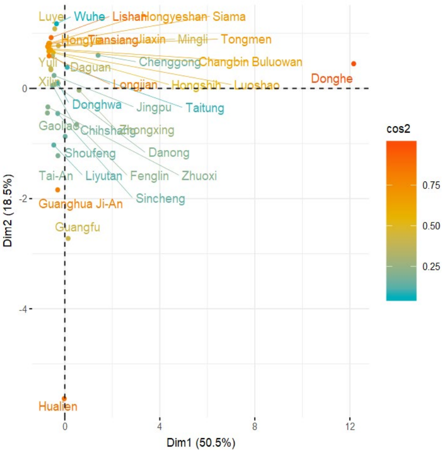


Figure 3. The New Coordinate System after PCA

Figure 3 displays the new coordinate system after the transformation. Similarly, the horizontal axis and the vertical axis of Figure 3 are PC1 and PC2, respectively. The dimensions are reduced from nine (hours) to two (PCs) for the precipitation data of 36 stations. Because PC1 and PC2 lie in the two directions with the first two greatest variances, the point that is the farthest from the origin indicates that the rainfall pattern of this station is much more distinct from that of the other stations. For each day available for PCA, we computed the Euclidean distance of each point to the origin and considered the largest distance. Given a cluster of a specified storm type, all these distances are obtained, and the threshold is determined using Equation (11) by setting p = 90th quantile. Thus, the criterion for detecting the anomalies is 10.19 (Table 3). The Donghe station is considered to have anomalies because its distance from the origin is 12.17, which exceeds the threshold of 10.185. The other stations are very close to the origin except for the Hualien station (the distance = 5.63). This plot implies that the variation of rainfall of the Donghe station is mainly captured by PC1, whereas that of the Hualien station is explained by PC2. In other words, among 36 stations, the temporal variation of the Donghe station is the largest.

The colors in Figure 3 indicate the quality of representation of individuals. cos2 equals to squared r in Equation (14). A high cos2 indicates a good representation of the individual by the PCs, and a low cos2 means that the individual is not perfectly represented by the PCs. From the color of the point Donghe station, we find that it is well represented by PC1. Moreover, the Hualien station is represented by PC2.

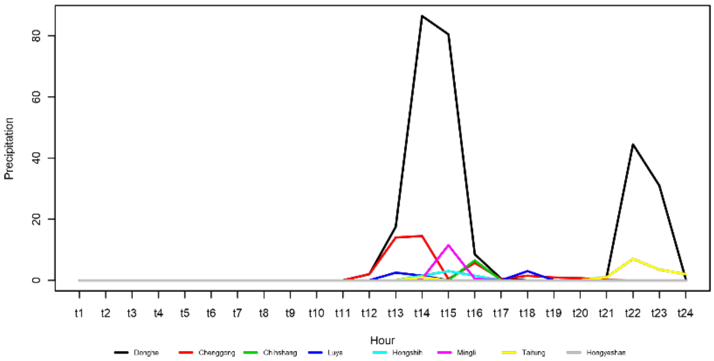


Figure 4. Rainfalls Observed by Donghe Station and Nearby Stations Donghe Station is detected to have anomalies. The neighboring stations are Chenggong, Chihshang, Luye, Hongshih, Mingli, Taitung, and Hongyeshan, from near to far.

Figure 4 shows the hourly rainfalls of the Donghe station and the other seven neighboring stations on Mar 27, 2020. It rained a lot at the Donghe station from 1 pm to 4 pm and from 10 pm to 12 am on this day (recorded rainfall: 86.5 mm at 2 pm and 80.5 mm at 3 pm). Hence, the temporal variation pattern in rainfall of the Donghe station is quite different from that of the other neighboring stations, classified as B1 (Table 2). Although our system identified that the Donghe station might have anomalies, further verification is needed to ensure whether anomalies exist.

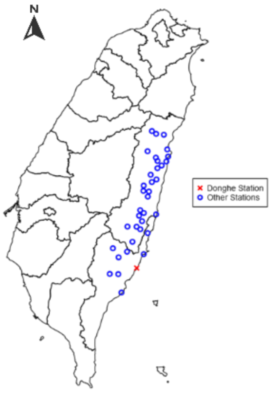


Figure 5. Anomaly Detected on Mar 27, 2020, in Cluster 4 of Frontal Rain

Figure 5 shows the detected result on March, 27, 2020. The red cross represents the Donghe station, and the blue circles represents the other stations in Cluster 4 of frontal rain. The red cross represents where Donghe Station is located, while there exist anomalies in the rain-falls that Donghe Station observed. The blue circles are other rain gauge stations, observing no anomalies, in Cluster 4.

Table 3. Threshold for Anomaly Detection of Frontal Rain

Cluster	1	2	3	4	5	6	7	8
Number of stations	61	27	50	36	55	47	15	6
Days Available	424	514	1401	947	551	682	592	1153
Threshold	8.93	6.86	12.04	10.19	10.39	7.64	6.09	5.23
Anomalies Detected	43	52	140	95	55	69	60	116
PAIVV	4	7	23	12	11	9	5	5

To save space, Tables 3 only presents the anomaly detection results of each cluster of frontal rain. “Days Available” shows the number of days that PCA can be performed. For each day, we computed the Euclidean distance from the subspace of PC1 and PC2. Then, we obtained the 90th quantile distance as the criterion for detecting anomalies. Taking Cluster 1 of the frontal rain type as an example, we calculated 424 maximum distances and set 8.93 (the 90th quantile of these distances) as the threshold. “Anomalies Detected” presents the number of anomalies (approximately one over ten of the available days) for each cluster. After our system discovered these anomalies, we examined them thoroughly and identified the possible anomalies for each cluster by visual verification (PAIVV).

Table 4. Nine Categories of Anomalies Detected by PCA for Each Storm Type

Code	B1	B2	B3	B4	B5	B6	B7	B8	B9	Sum
Frontal Rain	50	0	3	2	10	7	2	1	1	76
Meiyu	56	1	9	3	18	5	0	4	5	101
Convective Storms	87	1	4	0	4	0	1	0	4	101
Typhoons	20	1	0	0	0	0	0	0	1	22
Sum	213	3	16	5	32	12	3	5	11	300

According to Table 4, PAIVV of all storm types were divided into nine categories and are presented in Table 4. Categories B1, B2, B5, and B9 require satellite or weather radar images for anomaly verification. By contrast, anomalies belonging to categories B3, B4, B6, B7, and B8 were successfully identified. For each storm type, the number of B1 is the most, and the number of B5 is the second most.

5. Conclusion

We established an automated anomaly detection system for hourly precipitation data. Anomalies can occur at a station because of the marked difference in observed rainfalls between the specific station and other nearby stations. In other words, the rainfall observed by the station is extraordinarily higher or lower than that of the neighboring stations. The K-means cluster analysis is adopted to group the 297 stations based on geographical locations and rainfall characteristics as per the four primary storm types in Taiwan. Then, PCA is used to compute *d*, which is the Euclidean distance of the projected data point from the origin for each observation. When hours were taken as variables for PCA, *d* represented the temporal variation of the rainfall at each station in a specified cluster. When the value of *d* exceeded the threshold set, our system automatically indicates possible anomalies. The anomalies identified with PCA have nine categories. Some of them may not be anomalies, which still require additional verification. Hence, our system can effectively and efficiently screen out the potential anomalies to achieve the QC of hourly rainfall data.

The system is established using Shiny, a package developed by R Studio for users to create interactive web pages with R language. The URL of our online system is https://roam041.shinyapps.io/outlier_detection_v1/.

6. References

Boyle, J. S., and G. T. J. Chen, (1987). Synoptic aspects of the wintertime East Asian monsoon. *Monsoon Meteorology*, C. P. Chang and T. N. Krishnamurti, Eds., Oxford University Press, 125–160.

Branisavljević, N., Prodanović, D., Arsić, M., Simić, Z., & Borota, J. (2009). Hydro-Meteorological Data Quality Assurance and Improvement. *Journal of The Serbian Society for Computational Mechanics*, *3*(1), 228-249.

Chen, C., & Chen, Y. (2003). The Rainfall Characteristics of Taiwan. *Monthly Weather Review*, *131*(7), 1323-1341.

Chen, C., & Huang, J. (1999). A Numerical Study of Precipitation Characteristics over Taiwan Island during the Winter Season. *Meteorology and Atmospheric Physics*, *70*(3-4), 167-183.

Cox, D. (1957). Note on Grouping. *Journal of the American Statistical Association*, *52*(280), 543-547.

Fisher, W. (1958). On Grouping for Maximum Homogeneity. *Journal of the American Statistical Association*, *53*(284), 789-798.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6&7), 417-441 & 498-520.

Hudson, H., Mcmillan, D., & Pearson, C. (1999). Quality assurance in hydrological measurement. *Hydrological Sciences Journal*, *44*(5), 825-834.

Jolliffe, I. (2002). *Principal Component Analysis. 2nd ed* (2nd ed.). New York, NY: Springer-Verlag.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559-572.

Toe, M., Kanzaki, M., Lien, T., & Cheng, K. (2017). Spatial and temporal rainfall patterns in Central Dry Zone, Myanmar - A hydrological cross-scale analysis. *Terrestrial, Atmospheric and Oceanic Sciences*, *28*(3), 425-436.

Wang, S.-T., and H. Cheng, (1982). Natural seasons as shown by the variation of the circulation in Asia (in Chinese). *Atmos. Sci.*, 9, 125–146.

You, J., Hubbard, K., Nadarajah, S., & Kunkel, K. (2007). Performance of Quality Assurance Procedures on Daily Precipitation. *Journal of Atmospheric and Oceanic Technology*, *24*(5), 821-834.