Data Mining of Remotely-Sensed Rainfall for a Large-Scale Rain Gauge Network Design

Zhenzhen Liu, Huimin Wang, Jing Huang, and Lu Zhuo, *Member, IEEE*

***Abstract*—Rain gauges are able to measure point rainfall with high accuracy compared with remote sensing observations. However, a single rain gauge cannot provide continuous spatial coverage, and thus, rain gauge networks need to be designed in a way that will provide optimum rainfall information with a minimum number of gauges. While relatively inaccurate but long-term larger-scale satellite rainfall measurements are an ideal dataset to provide insight into local storm characteristics, such as rainfall patterns and local topographic influences on rainfall, these key characteristics would be useful in designing rain gauge networks. This article proposes a new scheme for a large-scale rain gauge network design that uses complementary data from satellite rainfall measurements. Principal component analysis (PCA), the elbow method, the intra-group sum of squares error and the gap statistic were used to calculate the optimum number of rain gauges respectively, while the locations of rain gauges were determined by cluster analysis with the influence of rainfall amount. The results show that PCA is the most effective method with multilevel variances, providing an optimum reference design number at 80% variances. In addition, the rainfall zones in Kenya had a close relationship with land cover, and more gauges will have to be deployed in the mountainous areas to reflect rainfall variations during the warm season in Kenya. The proposed scheme can also be used for other types of ground observation station designs in conjunction with remotely sensed hydrometeorological factors such as soil moisture, evapotranspiration and ice cover.**

***Index Terms*—Global precipitation measurement (GPM), network design, remotely sensed, satellite rainfall.**

I. Introduction

RAINFALL is one of the key components of Earth's water cycle, and as such, it has been investigated intensively in many scientific fields, such as meteorology, hydrology, ecosystem science, agriculture, and water resources. Development of rainfall measurement technologies has generally involved the following three phases: rain gauges, weather radar, and satellites. However, development of these technologies has not progressed at the same rate around the world. In most developing countries, such as Kenya, rain gauges are extremely sparse in many regions and weather radar is still absent; hence, researchers must rely on global satellite products to provide a general knowledge of the local rainfall. This fact should compel the community to reconsider the traditional development path of rainfall technologies and explore the possibility of developing rain gauge networks and weather radar with the aid of satellites.

The main advantages of using satellite-based remote sensing technologies are that one can scan large areas and compile millions of measurements of rainfall with relatively high spatial and temporal resolutions. However, uncertainties and biases in satellite rainfall data can arise because of the indirect measurement technique used, and this can lead to poor modeling behavior in many real hydrological and meteorological application [1]–[5]. Rain gauges, as one of the oldest and most common methods employed around the world, can measure point rainfall with relatively high accuracy compared with weather radar and satellites [6], [7]. However, rain gauges cannot provide rainfall data with continuous coverage in space, which is a requirement for most hydrological applications. A possible solution is to interpolate point measurements from a number of rain gauges into areal distributions of rainfall. The interpolation accuracy will depend on both the density of gauges and their spatial locations. Admittedly, we can increase the number of gauges to improve the quality of interpolated rainfall. However, it is expensive to operate a network with a large number of rain gauges. Moreover, a network of rain gauges that can survey the same area as weather radar (not to say satellite) with high spatial resolution would be practically impossible to set up and maintain. For this reason, there should be an optimum network design scheme that could capture adequate rainfall information with a minimum number of gauges. Given economic considerations, the limited rain gauges should fill in the most serious gaps from the perspective of water resources development and be employed at the optimal locations.

Because of the complexity and subjectivity of such issues, there is no universally agreed procedure in place for rain gauge network design, so rain gauge networks are generally developed in a haphazard manner [8]. Numerous technical "guidelines" or "considerations" exist for the deployment of a rain gauge network design, and these are based on information such as the nature of the catchment, its topographic influences, its drainage patterns, the accessibility and suitability of proposed locations, the costs of installing and maintaining the gauges, regional climate characteristics, and the purpose of the gauges [9]–[13]. More recently, some sophisticated techniques have been used

Manuscript received May 22, 2021; revised September 3, 2021 and November 6, 2021; accepted November 22, 2021. Date of publication November 26, 2021; date of current version December 13, 2021. This work was supported by the National Natural Science Foundation of China under Grants 91846203 and 42171081. *(Corresponding author: Jing Huang.)*

Zhenzhen Liu, Huimin Wang, and Jing Huang are with the State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering and Institute of Management Science Business School, Hohai University, Nanjing 210098, China (e-mail: liuzhenzhencs@163.com; hmwang@hhu.edu.cn; huangjingshow@hotmail.com).

Lu Zhuo is with the Department of Civil and Structural Engineering, University of Sheffield, S1 3JD Sheffield, U.K. (e-mail: lu.zhuo@sheffield.ac.uk).

Digital Object Identifier 10.1109/JSTARS.2021.3131157

to design rain gauge networks, such as kriging techniques, annealing algorithms, and information theory [14]–[19].

Compared with traditional methods that collect and excavate the limited rainfall-related information (such as rainfall spatial correlations and rainfall patterns) from inside or outside a study area, adoption of remotely sensed rainfall measurements for a rain gauge network design would obviously be a more direct and efficient way to approach data collection efforts. Rainfall information from weather radar has been used for rain gauge network design [8], [20]. However, weather radar tends to be also absent in most ungauged areas. Currently, we have already employed global satellite rainfall products for more than ten years (take the Tropical Rainfall Measuring Mission, TRMM, for example), but because of the poorer quality of such data, these rainfall products cannot replace rain gauge measurements at present, and rain gauges will still be the first choice for most applications in the near future. However, these relatively inaccurate but long-term larger-scaled rainfall measurements are an ideal dataset to provide insight into local storm characteristics, such as rainfall patterns and local topographic influences on rainfall, which are key characteristics that would be useful in rain gauge network design. Through analyzing long-term remotely sensed rainfall datasets, we could potentially deploy an "appropriate" number of rain gauges in "necessary" locations in a systematic manner.

For this reason, this article proposes a method for rain gauge network design based on data mining of remotely sensed rainfall using variable selection criteria. The design can be established easily and used to study redundancies in the designed rain gauge network. The model was implemented over the entire country of Kenya, and the results offer a rain gauge network solution that could serve as a reference in Kenya's future design of gauges. To the best of our knowledge, this is the first time that satellite rainfall information has been used for rain gauge network design work.

II. Study Area and Methods

A. Study Area and Data Sources

Kenya lies astride the equator between the longitudes 34°E and 42°E and latitudes 5°N and 4.5°S, where it covers an area of approximately 580 000 km^2. The country has a coastline on the Indian Ocean, and inland regions are diverse and include expansive terrain consisting of broad plains and numerous hills (see Fig. 1). The abundant geographical and climatic diversity provides an ideal study area for the analysis of rain gauge network design. The climate of Kenya varies by location, from mostly cool every day, to always warm/hot. Kenya has a warm and humid tropical climate on its Indian Ocean coastline, where rainfall and temperatures are higher throughout the year. At locations further inside Kenya, the climate becomes more arid.

A broad range of datasets covering geography, climate, and hydrology in Kenya were collected from various sources. Some sourcing information of these datasets is given in Table I. The elevation, annual rainfall, slope, and land cover with different spatial resolutions were mainly used for the following analysis, and these data are shown in Fig. 1. The remotely sensed rainfall from the global precipitation measurement (GPM) mission was used to drive the proposed method. The GPM, which is managed by the National Aeronautics and Space Administration's (NASA's) Goddard Space Flight Center, produces and distributes a wide variety of rainfall data products. All data from the mission are made freely available in a variety spatial and temporal resolutions on NASA's website.[1] The level 3 rainfall data with spatial and temporal resolutions of 0.1° and 1 day, respectively, were estimated by combining observations from all passive-microwave instruments in the GPM, and these data were used in this article.

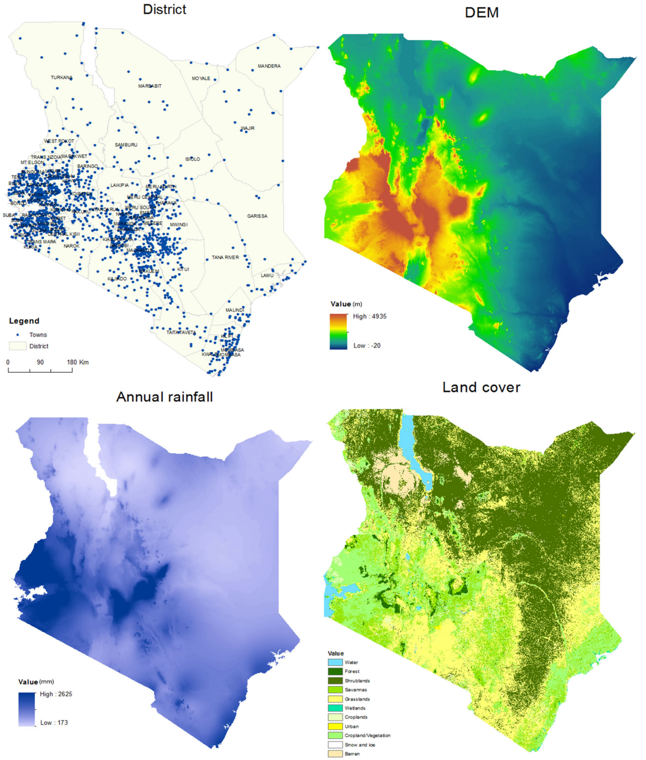


Fig. 1. Map of District, DEM, annual rainfall and land cover of Kenya.

Fig. 1. Map of District, DEM, annual rainfall, and land cover of Kenya.

TABLE I
Dataset Used in This Article

Datasets	Spatial resolution	Time coverage	Description
Satellite rainfall	0.10 degree	2014-2017	Global Precipitation Measurement (GPM)
Rain gauge rainfall	-	2014-2017	Kenya Meteorological Department
Ground weather data	-	-	Kenya Meteorological Department
Digital elevation model (DEM)	90 m	2014	Shuttle Radar Topography Mission (SRTM)
GIS data (e.g. districts, rivers, towns)	-	-	World Resources Institute
Land cover	0.50 degree	2012	Global Land Cover Project

[1] [Online]. Available: https://pmm.nasa.gov

TABLE II
INFORMATION OF RAIN GAUGE STATIONS IN KENYA

Station ID	Name	Begin Date	End Date	Elevation	Invalid Ratio
1	KEROCH	1973-07-19	2019-05-10	2184	36.36%
2	KITALE	1949-11-01	2019-05-10	1850.13	1.00%
3	EL DORET INTL	1996-01-01	2018-05-16	2150.06	11.40%
4	KISUMU	1949-09-09	2019-05-10	1208.53	36.20%

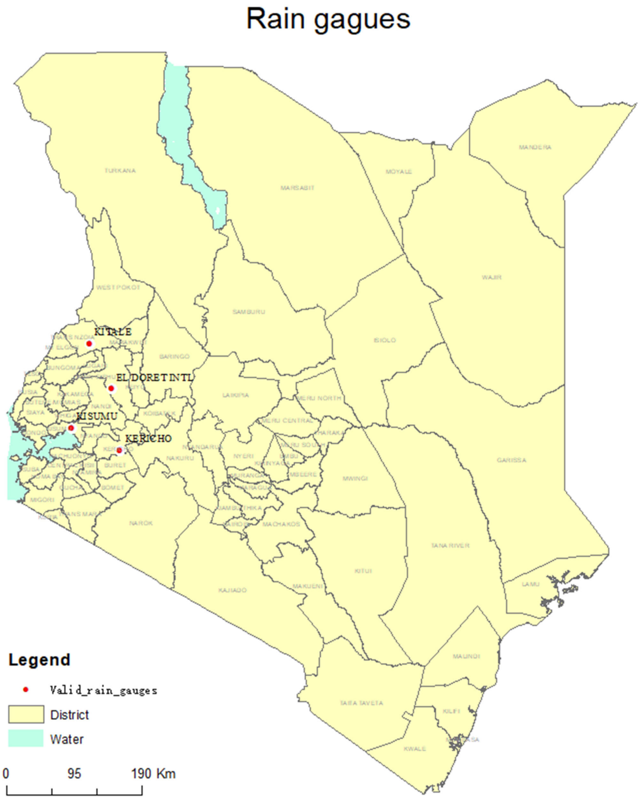


Fig. 2. Map of rain gauges in Kenya.

B. Satellite Data Evaluation

It is well known that satellite data have certain uncertainties, whose accuracy of rainfall measurement is less than that of rain gauges. In this article, the four-year (2014–2017) satellite data are used to capture rainfall variances over a continuous period of time rather than precise rainfall at a point. The core satellite measures rain using two science instruments: the GPM microwave imager and the dual-frequency precipitation radar. It is available for the satellite data within a certain error range to reconstruct a rain gauge network in Kenya. In order to assess the quality of satellite data, existed rain gauges in Kenya were regarded as a reference standard. Four rain gauges at one-day temporal scale with better continuity were selected to be valid criteria for satellite data, which displayed in red dots in Fig. 2. The relevant information of the valid rain gauges is given in Table II. Invalid Ratio means the proportion of invalid value.

Three verification indicators are used to evaluate the uncertainty of satellite data. The results of the evaluation of satellite are given in Table III. In general, the errors of satellite data with one-day scale are acceptable. It is worth remarking that satellite data in this article focus on the variance information of rainfall instead of the precise value.

TABLE III
VERIFICATION INDICATORS OF SATELLITE

Station ID	RMSE	PODy	ETS
1	0.6137	84.62%	87.50%
2	0.2801	60.04%	53.91%
3	0.2207	67.32%	64.52%
4	0.4072	80.42%	77.79%

1) Root Mean Square Error: Root-mean-square error (RMSE) is a common index in the error analysis. It calculates the error between rainfall measured by the existed rain gauges and the satellite. As shown in formula (1), R_{gauge} and R_{satelite} indicate gauge rainfall and the rainfall of the satellite grids which corresponding to valid gauges. *T* means time series

$$\text{RMSE} = \sqrt{\sum_{i=1}^{T} |R_{\text{gauge}} - R_{\text{satelite}}|}. \tag{1}$$

2) Probability of detection: Probability of Detection (POD): This parameter indicates the ability of the forecasting method to predict the probability of precipitation within a given threshold. PODy represents the probability of "yes" observations. As shown in formula (2), where hit is the number of correct positive forecasts, miss is the number of occurrences of the event which were not forecast. POD$y = 1$ indicates a perfect forecast, while a POD of 0 represents a poor forecast

$$\text{POD}y = \frac{\text{hit}}{(\text{hit} + \text{miss})}. \tag{2}$$

3) Equitable Threat score: Equitable threat score (ETS) is the index focused on the hit event. An ETS equal to 1 means a perfect forecast, while a value equal to 0 means the random or constant forecasts performed better. ETS is defined as follows, false alarm (Fa) is the number of incorrect positive forecasts, and correct negative (Cn) is the number of correct rejections

$$\text{ETS} = \frac{(\text{hit} - \text{Fa})}{(\text{hit} + \text{miss} + \text{Fa} - k)} \tag{3}$$

$$K = \frac{(\text{hit} + \text{Fa})(\text{Fa} + \text{miss})}{(hit + \text{Fa} + \text{miss} + \text{Cn})}. \tag{4}$$

C. Rain Gauge Network Design

The proposed scheme aims to determine the optimum combination of gauges for a rain gauge network design. It is desirable that rain gauges should be deployed in locations that can reflect the variability of rainfall. As no (or quite a limited number of) ground observation points exist in the study area, a key challenge involved finding the locations solely from satellite rainfall measurements that correspond to the variability of rainfall. The selected satellite grids from existing satellite grid networks were considered to be ideal places for deploying rain gauges, and these were named optimum grids (OGs). The center of each OG indicates one potential location for a rain gauge.

The appropriate numbers and locations of OGs are dependent upon the amount of original variance the network should retain. A group of OGs is considered to be the best combination of locations that can optimally provide the variance of the original satellite rainfall measurements within a given limited number. There are a number of data mining methods that can be used to choose a subset of the original variables, which approximate the retained variances. Cluster analysis (CA) was used to determine the optimum rain gauge locations in this article. CA is a technique for classifying a large amount of information into manageable and meaningful subsets of clusters according to a criterion that maximizes the similarity of cases within each cluster while maximizing the dissimilarity between groups that are initially unknown. One advantage of CA is that no prior knowledge is needed about which elements belong to which clusters. The k-means clustering technique was used in this article. K-means clustering treats each observation in the dataset as an object having a location in space; it finds a partition in which objects within each cluster are as close to each other as possible, and as far from objects in other clusters as possible.

If the background rainfall network has n satellite pixels, we have a dataset X of n variables, with p observations (e.g., daily rainfall in a year). The proposed scheme first allocates the original n satellite pixels to a given number (say, k) of clusters of variables. The challenge is to determine the value of k, which can be achieved by different ways. In this article, principal component analysis (PCA), the elbow method, the intragroup sum of squares error (SSE) and the gap statistic (GS) were applied to explore an optimal k value.

Second, steps are taken to define the centroid values of the clusters. The centroid for each cluster is the point at which the sum of distances from all objects in that cluster is minimized. There are distance measures that are Euclidean (i.e., can be measured with a "ruler") or attributes that are based on similarity. The Euclidean distance is the most straightforward and generally accepted way of computing distances between objects in a multidimensional space. Moreover, the rainfall connection among different satellite grids relies on their separated distances. So, we adopted this measure for the CA. Because of its gradient descent nature, the CA algorithm is sensitive to the initial placement of the cluster centers. To make the results stable, the initial satellite grids are designed to maximize their rainfall intervariability. The satellite grids are ordered by their long-term averaged rainfall, and the initial satellite grids are chosen evenly from this sequence. Then, each satellite grid is attributed to the closest cluster. The centroid position of each cluster is re-set to the mean of all satellite grids belonging to that cluster. This process is repeated until the data converge, which means that the centroid stays in a stable location to form a given number of clusters.

Finally, the satellite grid giving the medium averaged rainfall in each cluster is regarded as OG, which we named as CA-Med method. Actually, more reliable design scheme is supposed to take more social factors into account, including urban accessibility, population density and a series of other factors, which we will discuss in Section III. The main point of this article is to propose a method to reconstruct the network by using satellite data to capture the characteristics of rainfall information variance.

D. Determination of Optimal Gauge Number

The PCA is used to retain valuable satellite grids from the background rainfall measurements. In PCA, the $n \times p$ original dataset X is normalized to remove the dimensional impacts at first, and the $n \times n$ covariance matrix COV of the dataset is calculated. Correspondingly, the eigenvectors and eigenvalues of matrix COV are obtained. The $n \times n$ eigenvectors matrix E is multiplied with X to obtain n principal component Matrix P of the dataset. Moreover, sorting the contribution rate of variance, available principal components were selected when their accumulation is greater than a specific threshold.

The elbow method is used for further interpretations and to validate the consistency within the dataset designed to find the appropriate value k. This method estimates the percentage of variance explained as a function of the number of clusters and is used to construct the variance–number relationship curve. The optimum number of rain gauges is achieved by deciding on a threshold for the desired variance explained, and the components of lesser significance are ignored. For example, if we wish to maintain 90% of the variance found in the data and the number of clusters required to reach this percentage of explained variance is just 50 components (say, 1000 in total), then we can conclude that the background network is heavily redundant and 950 of the variables can be removed without significant loss of information. The final dataset will have significantly fewer dimensions than the original.

The intragroup SSE is a commonly method to estimate an optimal cluster counts in the group. It is acknowledged that the classification would be the best classification at minimum intragroup square sum error (SSE). The sum of squared error between each central pixel $\bar{x}$ and satellite pixel x in this group is calculated at different clustering counts. The term $\bar{x}$ indicates the rainfall of each central grids at time point T. Generally, original data recorded rainfall information for a period of time, and hence, the SSE should overlay the errors in all time points. The appropriate value k is a key turn when the slope (first-order derivative) of SSE under different clustering numbers, changes from steep to gentle. As shown in formula (5), N is the clustering numbers, M represents total pixel number of each cluster, and T means the dimension of time series of each pixel. In addition, the slope of SSE is directly presented through the first-derivative

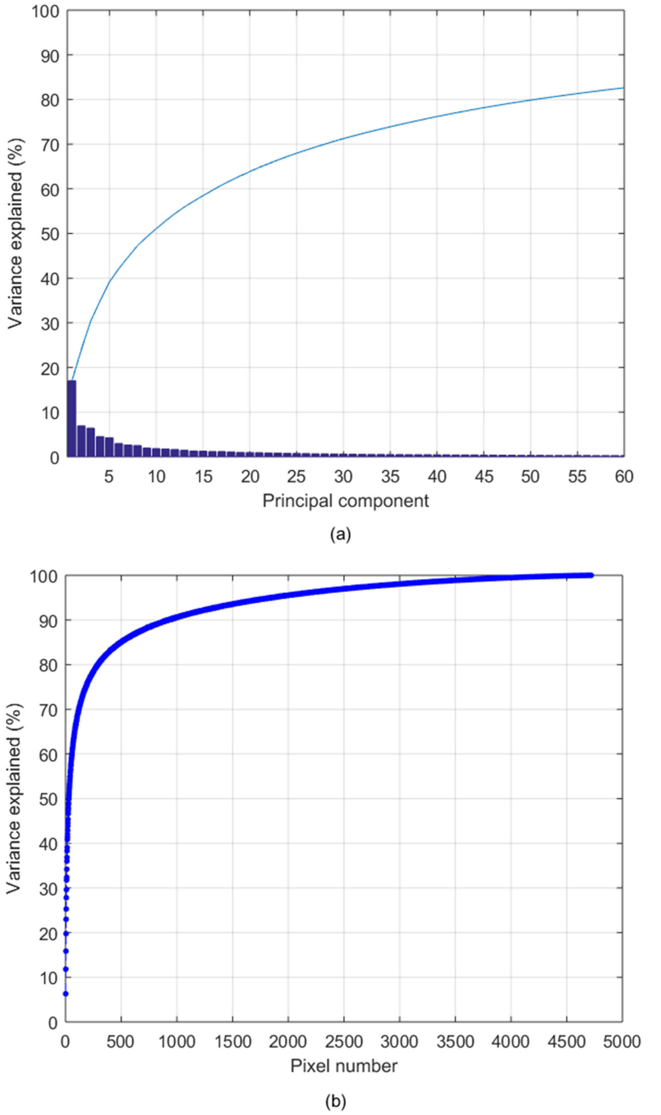


Fig. 3. Analysis of satellite rainfall in Kenya. (a) Principal component analysis. The bars mean variance explained of the principal component in descending order. The curve is the cumulative variance explained of the principal component. (b) Elbow method. The curve is the cumulative variance explained of the pixel number.

in formula (6)

$$\text{SSE} = \sum^{N}\sum^{M}\sum^{T}(x-\bar{x})^2 \tag{5}$$

$$\text{DSSE} = \frac{d\text{SSE}}{dN}. \tag{6}$$

The GS is a common method for confirming the number of groups in a set of data [21]. The technique combines with the K-means clustering algorithm to compare the alter in within-cluster dispersion to the expected one under an appropriate reference null distribution. The calculation of the GSs is divided into three steps as follows.

1) *Step 1:* Calculating W_k with clustering number k (k = *1, 2,…K*) for the observed satellite dataset. In k-means technique, we departed the dataset to k groups $C_1, C_2, \ldots, C_k$. m_r represents the total pixel number of C_r. Formula (7) defines the sum of distances between any two pixels (i, i') in C_r. It is worth noting that the distance mean in this article refers to the numerical difference rather than the spatial distance between pixels

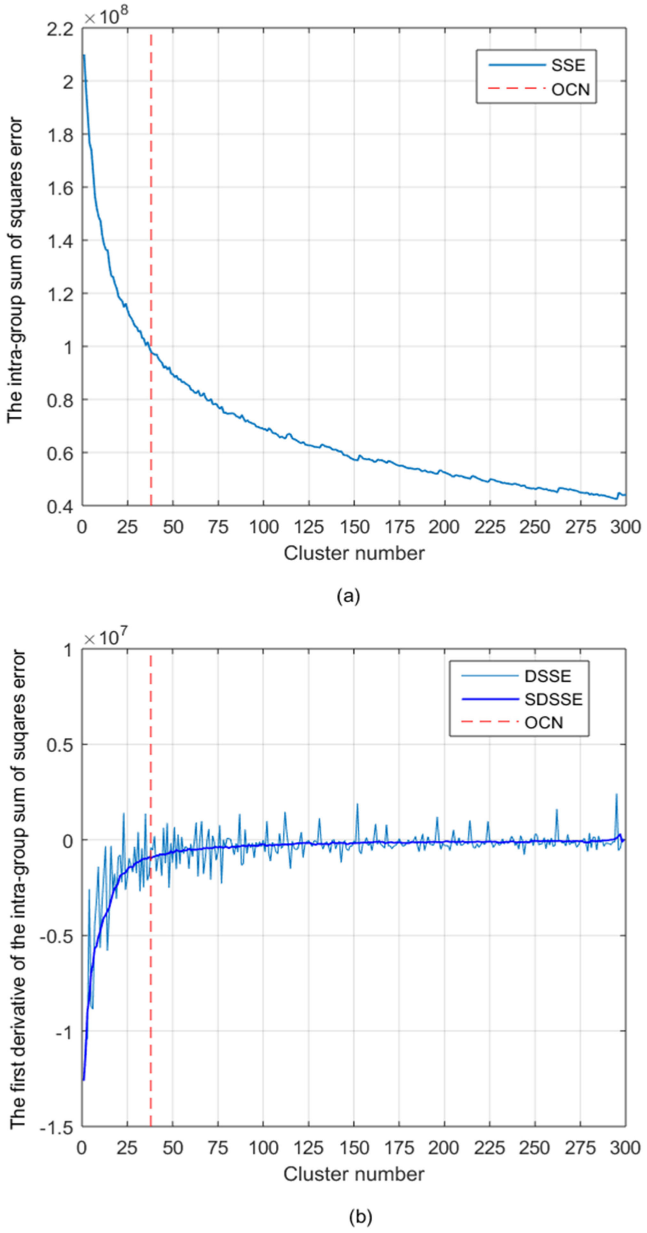
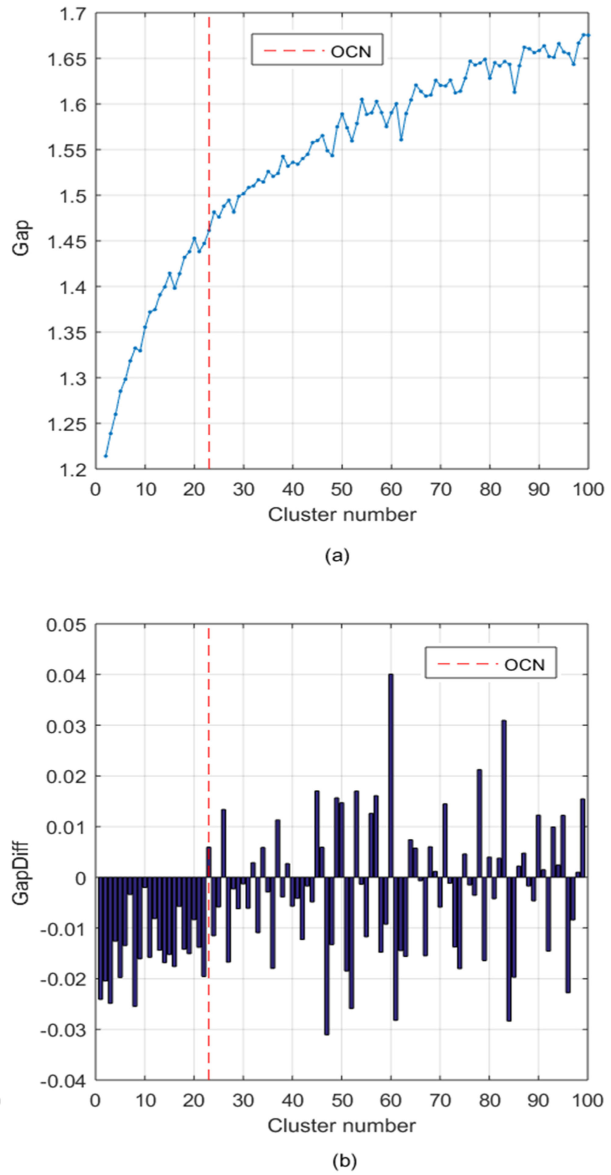


Fig. 4. Intragroup SSE analysis of satellite in Kenya. (a) SSE is the intragroup SSE. OCN is the optimal cluster number in the analysis. (b) DSSE is the first derivative of the intra-group SSE. SDSSE is the smooth first derivative of the intragroup SSE.

$$D_r = \sum_{i,i' \in C_r} dii' \tag{7}$$

$$W_k = \sum_{r=1}^{K} \frac{1}{2m_r} D_r. \tag{8}$$



(a)

(b)

Fig. 5. GS analysis of satellite in Kenya. (a) Gap is defined as formula (9). OCN is the optimal cluster number in the analysis. (b) GapDiff is defined as formula (10). OCN is the minimum value meeting the condition that the GapDiff was greater than 0.

2) *Step 2:* Setting up B reference datasets, and the W^*_{kb} of each reference dataset is calculated, respectively. In the process of calculation, considering the cost of time and computer memory, we adopt 1000 reference datasets to ensure the accuracy of results ($b = 12, \ldots, B$; $B = 1000$)

$$\text{Gap}\,(k) = \left(\frac{1}{B}\right)\sum_b \log(W^*_{kb}) - \log(W_k) \tag{9}$$

3) *Step 3:* The standard deviation $\text{sd}(k)$ of the entire reference dataset was obtained, and then the GapDiff was calculated as the formula (11). The optimal k was the minimum value meeting the condition that the GapDiff was greater than 0

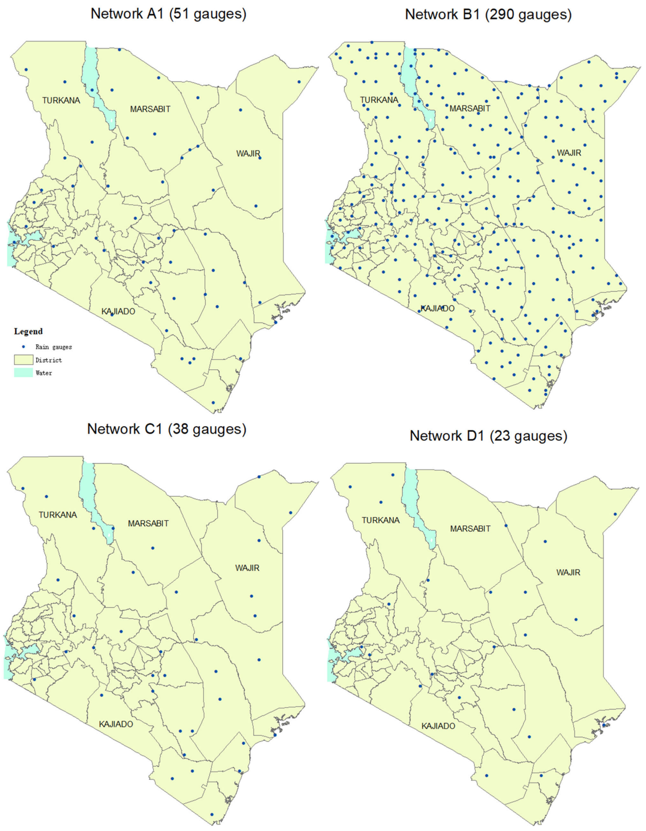


Fig. 6. Designed rain gauge networks for Kenya based on CA-Med. Network A was obtained by the PCA technique, B was obtained by the elbow method, C was obtained by the SSE method, and D is the output from the GS technique.

$$\text{sd}\ (k) = \sqrt{\left(\frac{1}{B}\right)\sum_b \left(\log(W^*_{kb}) - \left(\frac{1}{B}\right)\sum_{b=1}^{B}\log\left(W^*_{kb}\right)\right)^2} \tag{10}$$

$$\text{GapDiff}\ (k) = \text{Gap}\,(k) - \text{Gap}\,(k+1) + \text{sd}\,(k)\sqrt{1+\frac{1}{B}}. \tag{11}$$

E. Validation Methods

Performance evaluation of the rain gauge network design is a challenge because no standard assessment criteria exist to indicate what kind of network is the most appropriate one for the given study area. More importantly, the study area in Kenya that is in need of a rain gauge network design lacked rainfall information, which made validation work difficult.

In essence, the designed network should be an effective network, which means the network should contain maximum information at the cost of a minimal number of gauges. In other words, the chosen OGs should maintain the dominating rainfall information of the original satellite grid network, and further deployment of OGs should not significantly increase the amount of rainfall information. For this reason, we first evaluated the rainfall information of optimum satellite grids and investigated the relationship between the selected number of OGs and variance explained. In the second stage of validation, two indicators [the Pearson correlation coefficient and Nash–Sutcliffe (NS) coefficient] that can estimate the rainfall discrepancy of the designed network and the original network were introduced. The Pearson correlation coefficient (r) can estimate the systematic deviation between the OG's rainfall (R_m) and the original satellite grid's rainfall (R_o), and it is written as follows:

$$r_{R_0,R_m} = \frac{E\left[R_m R_0\right] - E\left[R_m\right] E\left[R_0\right]}{\sqrt{\left(E\left[{R_m}^2\right] - E[R_m]^2\right) \times (E\left[{R_0}^2\right] - E[R_0]^2}} \tag{12}$$

where E represents the mean value of the corresponding vector. The NS coefficient [22] is generally used to assess how well hydrological models predict events. Here, the rainfall of the designed network and existing network were regarded as modeled and observed values, respectively. The NS coefficient is calculated as follows:

$$\text{NS} \; = \; 1 - \frac{\sum \left(R_0^t - R_m^t\right)^2}{\sum \left(R_0^t - E\left[R_0\right]\right)^2} \tag{13}$$

where the superscript t refers to the time-step of the storm and NS$\in$[1,-∞). The closer NS is to 1, the more accurate the designed scheme is.

In addition, to reveal the physical meanings behind the proposed statistical scheme, the relationships between the designed rain gauge locations and physical factors were investigated. For example, the possible influences of local climate and geography on the rain gauge network design could be related to the rainfall regime, digital elevation model (DEM) data, slope, and land cover. Thus, we evaluated whether the designed network satisfied the basic tendencies of these relationships.

III. Results and Discussion

A. Optimal Gauge Number in the Network

1) Variance–Number Relationship Analysis: With 4721 satellite grids located within Kenya, redundancy of rainfall should exist in the satellite grid network. The PCA was applied to the rain gauge network to provide a measurement of the optimum satellite grids and calculate acceptable losses for the total information. The principal components of satellite rainfall are shown in Fig. 3(a). The first principal component carried about 18% of the total variance, while the top ten components brought this value up to around 50% of the total variance. To better show the relationship between the principal component numbers with the variance explained, thresholds of the desired variance explained were set to 70%, 75%, 80%, 85%, 90%, 95%, 97% and 98%. The required number of components and pixel number are given in Table IV. For the principal component, it can be seen that 29 components were sufficient to retain 70% of the information. When the variance threshold was set to 80%, 51 variables contained the required information, thus indicating that there was a relatively high level of redundancy in the satellite data.

TABLE IV
Number of Components and Pixels to Reach % Variance Threshold for Kenya

Variance	Components	Pixels
70.0	29	121
75.0	38	183
80.0	51	290
85.0	72	493
90.0	108	914
95.0	187	1843
97.0	256	2497
98.0	316	2961

The pattern can also be revealed by the elbow curve (also known as the variance–number curve), which is shown in Fig. 3(b). With the increase of pixel number, the variance explained grew quickly when it was smaller than 0.8. The increase rate slowed down obviously after this point. The ratio was further weakened when the variance approached 0.9. For this reason, we supposed that the turning point was located in the domain between 0.7 and 0.9. Obviously, the approximate number corresponding to 80% variance explained was 290, which was far from those obtained by the PCA. The magnitude gap between two methods would be discussed in Section III. A value of 20% information lost should be acceptable in most situations as large numbers of insignificant variables can be discarded. Both the PCA and elbow method results showed that the curve gradually became flat when the variance rate increased to 80%. Therefore, to explore the rain gauge network design under different data mining situations, the variance threshold was set to 80% for the further investigation. The exact value chosen for the variance should depend on the purpose of gauge use. Other factors such as the topographic influences and the costs of installing and maintaining the gauges could also be considered.

2) Deviation-Number Relationship Analysis: Both of the above methods determine the optimal k value of rain gauges by screening the variance contribution rate. However, the intragroup SSE and the GS estimate the relationship between clustering deviation and the clustering numbers to explore optimal k value. As the clustering number increased, SSE decreased rapidly, and the slope of SSE (DSSE) flattened rapidly in Fig. 4. The SDSSE means the smoothing curve of DSSE. Obviously, a demarcation line to distinguish SSE from steep to gentle was obtained as an optimal clustering number (OCN), which was between 25 and 50 in this article. Admittedly, SSE only provide a range rather than an exact unique value, and hence, we determined the unique k of SSE by replacing with the medium. It was a simple way to judge that when the number of clusters is greater than 38, the DSSE was hardly vary and also close to 0. However, the GS displayed a diverse output in Fig. 5. When cluster number was over 23, the GapDiff was larger than 0. In other words, the minimum group number was 23 to design a network with K-means clustering in Kenya.

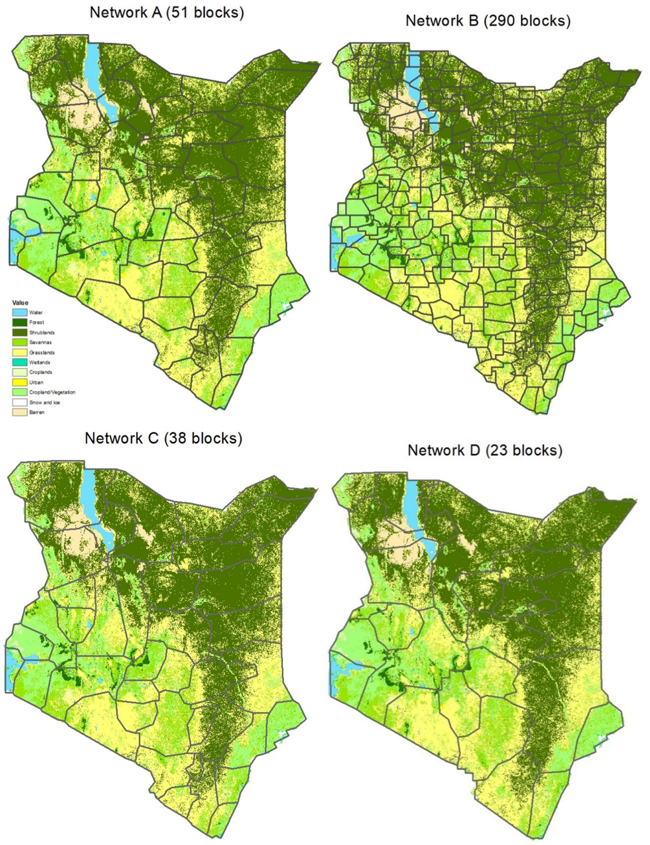


Fig. 7. Maps of derived rainfall zones with land cover as background.

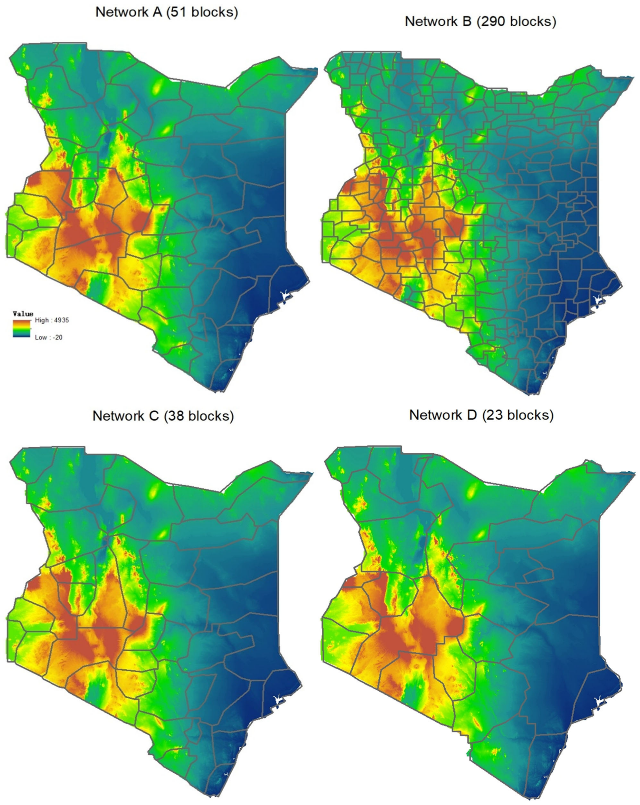


Fig. 8. Maps of derived rainfall zones with DEM as background.

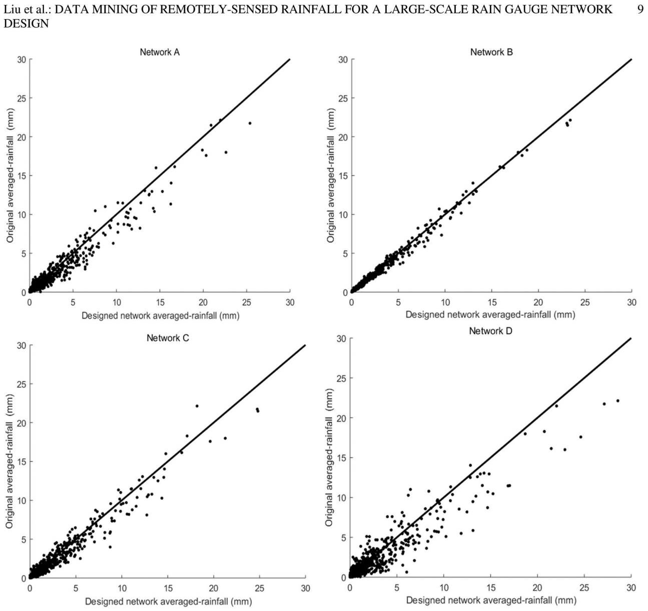


Fig. 9. Correlation between averaged rainfall of all the GPM pixels based on CA-Med in Kenya.

B. Rain Gauge Network Design Using CA-Med Methods

Based on the designed rainfall zones and the proposed selection criteria, the optimum locations of rain gauges for Kenya were determined, and these locations are shown in Fig. 6. These points in the figure (named OGs in this article) give valuable insight into the preferential areas for rain gauge locations. Network A is obtained by PCA technique, B is found by Elbow method, C is obtained by SSE method, and D is the outputs of GS technique. To better show the locations of gauges, the district boundaries of Kenya were also depicted. For network A, one can observe a slight change of gauge density among different districts. For example, the areas of Marsabit county and Wajir county were quite similar (the names of the districts are shown in Fig. 1), but the designed gauge numbers were quite different (8 compared with 3). There is also a connection between the optimum locations of rain gauges in different networks. In networks A and C, the number of rain gauges was all the same in Turkana, Wajir, and Kajiado. However, with less rain gauges in network D, some western cities no longer need to install devices. Therefore, it can be concluded that the distribution of rain gauges in different districts is obviously different with less devices (networks A, C, and D), and the gauges in western region is significantly reduced compared with the existing rain gauge network. While the network is more intensive (network B), the 290 gauges were quite evenly distributed across the country. Except the network B1, the distribution of rain gauges among different networks is quite similar. There are even some locations that are almost the same. The network B1 can retain a larger variance, but at the cost of many more rain gauges. The exact value chosen for the variance are suggested to depend on the purpose of gauge use.

C. Relationship Between Rainfall Zones With Land Cover and Terrain

The rainfall zones indicate the maximum similarity within zones and discrepancy between zones, which should have a close relationship with the characteristics of the local climate and geography. Given economic considerations, the limited rain gauges should be representative in a given region. In other words, they need to represent their surrounding area, which becomes more complex with lower gauges numbers and larger rainfall zones. As the proposed scheme is purely statistical, to reveal the physical causes behind the derived rainfall zones, the comparisons of rainfall zones and land cover as well as terrain are shown in Fig. 7 and Fig. 8. From the four figures in Fig. 7, it can be seen that land cover played a key role in determining the rainfall zones. One can observe that the boundary of rainfall zones to some extent followed the boundary of different land cover types. For example, the southeastern part of Kenya is covered by shrublands, grasslands, and a cropland/natural vegetation mosaic. This is logical based on ecohydrology because rainfall has a major influence on land cover types. Despite some crossover, most land covers in each of the grids of the network were consistent (see networks C and D in Fig. 7). Admittedly, there were some areas in rainfall zones that violated the boundaries of two land cover types, but this makes sense as the formation of rainfall is very complicated and is determined by a variety of factors. Many traditional rain gauge network design methods have attempted to find the patterns among the possible influencing factors, but this approach has proven to be inefficient and impractical. Conversely, long-term satellite rainfall observations provide an efficient way to examine the direct rainfall characteristics and delineate rainfall zones.

In terms of terrain, there was also connection between it and the rainfall zones, and the relationships were a bit stronger than the ones between the land cover and rainfall zones. Take the west valley part, for example, where it can be seen from network D that the boundaries of rainfall zones generally followed the variation of terrain. For network B, similar patterns can be seen, although the numerous tiny blocks bring about more uncertainty. In summary, land cover was found to have significant links with the rainfall zones in Kenya, and the terrain effect should not be neglected either.

D. Comparison and Validation of Design Methods

The method of optimum rain gauge network design proposed in this article has two crucial steps: the first step is to determine how many rain gauges are placed, and the second step has to solve the problem of rain gauge locations. For the first one, four methods were all applied to explore a most appropriate pattern. Moreover, the medium rainfall (CA-Med) is also validated to select optimum locations of rain gauge.

With the designed numbers corresponding to 80% variances, the PCA (network A) and the elbow methods (network B) applied 51 and 290 gauges to design the network. Moreover, the SSE (network C) and the GS (network D) utilize 38 and 23 gauges to design the network according to the above analysis. Admittedly, four methods of determining the optimum design number showed an inconsistent outcome on the base of the above, but they are all interpretable. The design number obtained by the elbow method was quite different from others, it calculated variance directly in units of each pixel, unlike PCA, which replaced numerous pixels with principal components to reducing the design numbers. There is no doubt that the outcome of the elbow method is linked with the resolution of satellite pixels, higher resolution is corresponds to more pixels.

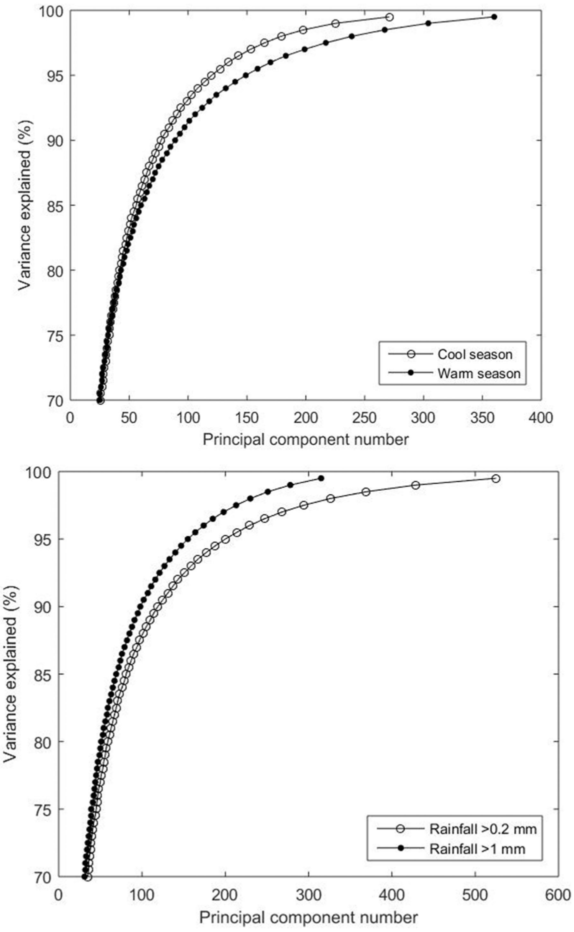


Fig. 10. Comparison of variance-number relationships under season (top) and rainfall intensity (bottom).

To further evaluate the efficiency of the four proposed scheme, the rainfall data derived by the designed network were compared to the completed satellite rainfall observations. The Pearson correlation coefficient and Nash coefficient were used to quantify the relationships between the two rainfall sets. The results are shown in Fig. 9 for networks A–D. The detailed statistics are given in Table V. In Fig. 9, black dots represent rainfall pairs of the original and designed networks for one time-step, with their fitted linear relationship shown in lines. The linear relationship between them fit the data quite well. Overall, four networks produced good estimates for the average rainfall. This can be given in Table V, which gives the corresponding Pearson correlation coefficient for each network. All the correlation coefficients were larger than 0.95. In addition, some scholars suggested a threshold value of NS coefficient between 0.5 and

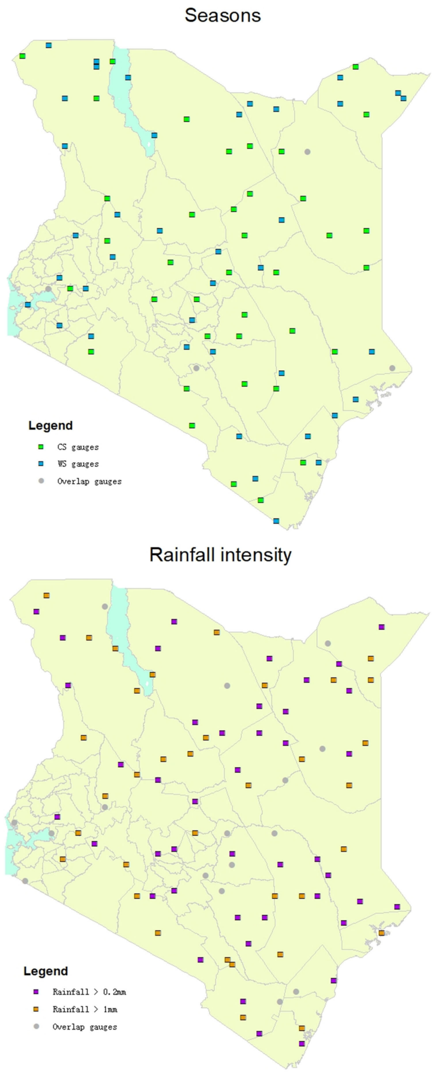


Fig. 11. Designed rain gauge networks for different seasons (top) and different rainfall thresholds (bottom).

TABLE V
CORRELATION COEFFICIENTS AND NS COEFFICIENTS OF THE DESIGN NETWORKS BASED ON CA-MED

Network	Correlation	Nash-Sutcliffe
A	0.98	0.92
B	0.99	0.99
C	0.95	0.85
D	0.95	0.82

0.65 could imply a good prediction ability [23], [24]. We can see NS coefficients of four schemes all indicated a satisfactory predictive ability in Table V.

There are significant connections between the PCA and the others, in short, the results of the other three methods could be explained by the PCA corresponding multi-level variances. For example, the design number obtained by PCA corresponding to 70% variances (29 gauges) was essentially in agreement with the that obtained by GS (23 gauges). Moreover, with the increase of variances to 75%, the design number obtained by PCA (38 gauges) was equal to that of SSE. Even if the result was greatly difference obtained by the elbow method (290 gauges), it also could be approximately related to the PCA corresponding 98% variances. Therefore, the design numbers of four methods are not inconsistent, and the PCA was the most available method with multi-level variances. In addition, it highlights that the construction and maintenance costs of the rain gauge network should be reduced as much as possible, besides the accuracy and redundancy. Specifically, the Pearson correlation coefficient and Nash coefficient (see Fig. 9) all displayed a best performance, but it is a considerable expense to maintain 290 rain gauges. Therefore, in the case of sufficient economic support, network B was no doubt the best scheme, while in the limited economic conditions, the choice of network A was the best practical scheme. The Correlation and NS coefficient of network A was over 0.9 (see Table V), but the total numbers of rain gauges was only 1/5 of that of network B. Overall, the PCA with 80% variances was undoubtedly the most effective method to obtain an optimum design numbers.

Although other methods can also be interpreted by adjusting the variance as explained (the elbow method), different choices are needed for the demarcation point (the SSE method). The PCA has more prominent advantages to carry out on multilevel variances. The CA-Med is a way of selecting the representative OG, which has been validated. In Fig. 9, the Pearson correlation coefficient-based CA-Med design method shows a good linear relationship with the average rainfall of the whole satellite network. Moreover, it can be clearly seen that the Pearson correlation coefficients and Nash coefficients are all greater than 0.95 and 0.8, which proves the effectiveness of design scheme of the CA-Med.

E. *Sensitivity Analysis of the Designed Network*

Given the attendant complexities, there are still some possible uncertainties associated with the designed rain gauge network. The comparisons of variance–number curves between different seasons and rainfall thresholds are shown in Fig. 10 (bottom). It can be seen from the figure that the required numbers were almost the same for the two seasons when the principal component was less than 60 (corresponding to about 85% of the variance). The discrepancy between the two curves increased from this point, it shows that more gauges were required with more than 85% variances in warm season (WS). The variance–number curves for the area-averaged thresholds of 0.2 and 1 mm are drawn in Fig. 10 (top). It can be observed from this figure that the curves were almost the same when the principal component was less than 50 (corresponding to about 80% of the variance), but more gauges should be placed in mild rain conditions when the variances was more than 80%. The rainfall threshold (e.g., 0.2 mm) means that only rainfall values exceeding the 0.2 mm threshold were used for the network design. A value of 0.2 was chosen instead of 0 to avoid small, negligible values. We only considered the average value in this article. In fact, the spatial and temporal variability of rainfall may also affect the network design. However, we could not feasibly enumerate and analyze all the possible uncertainties in this article.

The designed networks (80% variance) for different situations are shown in Fig. 11. The overlap OGs refer to the satellite grids that existed in both networks. With an acceptable tolerance, the adjacent OGs were considered to be overlap OGs as well. The overlap OGs are shown as gray solid circles in Fig. 11. After excluding the overlap OGs, the remaining ones were plotted in solid boxes. As is shown in Fig. 11, the numbers of OGs for the overlap, WS, and cool season were 4, 39, and 38, respectively. In other words, 9% of the OGs in the warm network and 10% of the OGs in the cool network overlapped. In terms of the rainfall threshold, the numbers of OGs for the overlap, 0.2 mm rainfall threshold, and 1 mm rainfall threshold were 16, 43, and 35, respectively. The same OGs accounted for 27% and 31% of the two networks. There was no obvious pattern in the distribution of the remaining OGs, although slight deviations of the OG density existed in some regions.

IV. Conclusion

The results of this article show that satellite rainfall measurements combined with selection criteria are an effective tool for the design of large-scaled rain gauge networks. Moreover, this new methodology can be theoretically used in all ungauged catchments as it only requires rainfall data provided by remote sensing datasets with global coverage. The design numbers derived from the PCA, the elbow method, the intragroup SSE and GS were applied to determine the design rain gauge numbers, while they do not represent physical rain gauges, and therefore, criteria selection methods are required to identify the optimum rain gauge locations. Based on CA, the background variables are classified by a given number of clusters, which are named rainfall zones. The satellite grids with the medium sum of rainfall in each rainfall zone are chosen as the OGs, respectively. The correlation of the area-average rainfall between the designed network and the original satellite grids was computed to test the proposed schemes. The results followed the basic patterns of local climate and geographical conditions and proved the rationality of the proposed schemes.

In summary, the following five main conclusions can be drawn from this article.

1) PCA is an effective method to assess the optimum satellite grids of satellite rainfall measurements.
2) CA using remotely sensed rainfall observations is an efficient and practical method for larger-scaled rain gauge deployment.
3) Method based on medium sum of rainfall is a practical way to search optimal grids in CA.
4) Rainfall zones in Kenya have a close relationship with land cover, and the elevation is also a considerable factor.
5) Rain gauge network design is season-dependent and intensity-dependent.

Since this is the first time that satellite rainfall has been fully used in rain gauge network design, we hope this article will stimulate more studies by the hydrological community so that a wide range of regions and climate conditions can be investigated and a clear pattern can be established. The network design method mainly focuses on the natural environment, some social factors are also supposed to be considered, such as urban density, traffic accessibility, or populations. It should be pointed out that the remotely sensed rainfall data are subjected to many uncertainty sources [25]–[28]. The possible influence of the satellite rainfall uncertainty on the rain gauge network design will be investigated in future work. In other places, such as the U.K., where dense radar networks can provide real-time rainfall with high spatial and temporal resolutions and satellites can scan the whole area by taking millions of measurements, the main challenge will be to determine what are the best gauges to shut down given the high costs of operating and maintaining rain gauges. In fact, a marked decline of hydrometric network density in many parts of the world has been noted, which is possibly being driven in part by reasons, such as insufficient funding and inadequate institutional frameworks [11], [29], [30]. In an effort to address this issue, an efficient and feasible scheme for streamlining rain gauge networks is needed. In addition to its utility in constructing rainfall gauge networks, the proposed method would also be applicable to the streamlining issue discussed above. How to determine the best combination of rain gauges, weather radar, and satellites to offer optimum rainfall information with low operating costs will be investigated in future studies.

References

[1] M. J. M. Cheema and W. G. Bastiaanssen, “Local calibration of remotely sensed rainfall from the TRMM satellite for different periods and spatial scales in the Indus Basin,” *Int. J. Remote Sens.*, vol. 33, no. 8, pp. 2603–2627, 2012.

[2] A. S. Gebregiorgis and F. Hossain, “Understanding the dependence of satellite rainfall uncertainty on topography and climate for hydrologic model simulation,” *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 704–718, Jan. 2013.

[3] V. Maggioni, P. C. Meyers, and M. D. Robinson, “A review of merged high-resolution satellite precipitation product accuracy during the tropical rainfall measuring mission (TRMM) era,” *J. Hydrometeorol.*, vol. 17, no. 4, pp. 1101–1117, 2016.

[4] E. I. Nikolopoulos *et al.*, “Understanding the scale relationships of uncertainty propagation of satellite rainfall through a distributed hydrologic model,” *J. Hydrometeorol.*, vol. 11, no. 2, pp. 520–532, 2010.

[5] C. J. Skinner *et al.*, “Hydrological modelling using ensemble satellite rainfall estimates in a sparsely gauged river basin: The need for whole-ensemble calibration,” *J. Hydrol.*, vol. 522, pp. 110–122, 2015.

[6] P. Salio, M. P. Hobouchian, Y. G. Skabar, and D. Vila, “Evaluation of high-resolution satellite precipitation estimates over southern South America using a dense rain gauge network,” *Atmos. Res.*, vol. 163, pp. 146–161, 2015.

[7] R. Xu *et al.*, “Ground validation of GPM IMERG and TRMM 3B42V7 rainfall products over southern Tibetan Plateau based on a high-density rain gauge network,” *J. Geophys. Res., Atmos.*, vol. 122, no. 2, pp. 910–924, 2017.

[8] Q. Dai *et al.*, “A scheme for rain gauge network design based on remotely sensed rainfall measurements,” *J. Hydrometeorol.*, vol. 18, no. 2, pp. 363–379, 2016.

[9] M. Al-Zahrani and T. Husain, “An algorithm for designing a precipitation network in the south-western region of Saudi Arabia,” *J. Hydrol.*, vol. 205, no. 3, pp. 205–216, 1998.

[10] T. Husain, “Hydrologic uncertainty measure and network design,” *JAWRA J. Amer. Water Resour. Assoc.*, vol. 25, no. 3, pp. 527–534, 1989.

[11] A. K. Mishra and P. Coulibaly, “Developments in hydrometric network design: A review,” *Rev. Geophys.*, vol. 47, no. 2, pp. 2415–2440, 2009.

[12] R. Moore, D. Jones, D. Cox, and V. Isham, “Design of the HYREX raingauge network,” *Hydrol. Earth Syst. Sci. Discuss.*, vol. 4, no. 4, pp. 521–530, 2000.

[13] M. L. Morrissey, J. A. Maliekal, J. S. Greene, and J. Wang, “The uncertainty of simple spatial averages using rain gauge networks,” *Water Resour. Res.*, vol. 31, no. 8, pp. 2011–2017, 1995.

[14] S. K. Adhikary, A. G. Yilmaz, and N. Muttil, “Optimal design of rain gauge network in the Middle Yarra River catchment, Australia,” *Hydrol. Processes*, vol. 29, no. 11, pp. 2582–2599, 2015.

[15] Y. C. Chen, C. Wei, and H. C. Yeh, “Rainfall network design using kriging and entropy,” *Hydrol. Processes*, vol. 22, no. 3, pp. 340–346, 2008.

[16] M. Nour, D. Smit, and M. G. El-Din, “Geostatistical mapping of precipitation: Implications for rain gauge network design,” *Water Sci. Technol.*, vol. 53, no. 10, pp. 101–110, 2006.

[17] D. Tsintikidis *et al.*, “Precipitation uncertainty and raingauge network design within Folsom Lake watershed,” *J. Hydrol. Eng.*, vol. 7, no. 2, pp. 175–184, 2002.

[18] T. H. Volkmann, S. W. Lyon, H. V. Gupta, and P. A. Troch, “Multicriteria design of rain gauge networks for flash flood prediction in semiarid catchments with complex terrain,” *Water Resour. Res.*, vol. 46, no. 11, 2010, Art. no. W11554.

[19] H. C. Yeh, Y. C. Chen, C. Wei, and R. H. Chen, “Entropy and kriging approach to rainfall network design,” *Paddy Water Environ.*, vol. 9, no. 3, pp. 343–355, 2011.

[20] A. A. Bradley *et al.*, “Raingage network design using NEXRAD precipitation estimates,” *JAWRA J. Amer. Water Resour. Assoc.*, vol. 38, no. 5, pp. 1393–1407, 2002.

[21] R. Tibshirani and W. T. Hastie, “Estimating the number of clusters in a data set via the gap statistic,” *J. Roy. Stat. Soc. Ser. B (Stat. Methodol.)*, vol. 63, no. 2, pp. 411–423, 2001.

[22] J. Nash and J. V. Sutcliffe, “River flow forecasting through conceptual models Part I—A discussion of principles,” *J. Hydrol.*, vol. 10, no. 3, pp. 282–290, 1970.

[23] D. N. Moriasi *et al.*, “Model evaluation guidelines for systematic quantification of accuracy in watershed simulations,” *Trans. ASABE*, vol. 50, no. 3, pp. 885–900, 2007.

[24] A. Ritter and R. Muñoz-Carpena, “Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments,” *J. Hydrol.*, vol. 480, pp. 33–45, 2013.

[25] Q. Dai, D. Han, M. A. Rico-Ramirez, and T. Islam, “Modelling radar-rainfall estimation uncertainties using elliptical and archimedean copulas with different marginal distributions,” *Hydrol. Sci. J.*, vol. 59, no. 11, pp. 1992–2008, 2014.

[26] Q. Dai *et al.*, “Radar rainfall uncertainty modelling influenced by wind,” *Hydrol. Processes*, vol. 29, no. 7, pp. 1704–1716, 2015.

[27] B. Zhao *et al.*, “Assessing the potential of different satellite soil moisture products in landslide hazard assessment,” *Remote Sens. Environ.*, vol. 264, 2021, Art. no. 112583.

[28] Q. Dai *et al.*, “Adjustment of radar-gauge rainfall discrepancy due to raindrop drift and evaporation using the weather research and forecasting model and dual-polarization radar,” *Water Resour. Res.*, vol. 55, no. 11, pp. 9211–9233, 2019.

[29] J. C. Rodda, “Guessing or assessing the world’s water resources?,” *Water Environ. J.*, vol. 9, no. 4, pp. 360–368, 1995.

[30] J. C. Rodda *et al.*, “Towards a world hydrological cycle observing system,” *Hydrol. Sci. J./J. Des Sci. Hydrol.*, vol. 38, no. 5, pp. 373–378, 1993.

Zhenzhen Liu was born in Hunan, China, in 1986. She received the B.E. degree in hydrology and water resources engineering from the Changsha University of Science and Technology, Changsha, China, in 2009, the M.S. degree in hydrology and water resources from Sun Yat-sen University, Guangzhou, China, in 2011. She is currently working toward the Ph.D. degree in management science and engineering with Hohai University, Nanjing, China.

Her research interests include hydroinformatics, real-time flood forecasting and risk assessment, and management of flood.

Huimin Wang was born in Shanxi, China, in 1963. She received the B.S. degree in mathematics from Shanxi University, Shanxi, China, in 1984, the M.S. degree in mathematics from Chinese Academy of Science, Beijing, China, in 1989, and the Ph.D. degree in management science and engineering from China University of Mining and Technology, Xuzhou, China, in 1997.

She is currently a Professor with the Department of Management Science and Information Management, School of Business, Hohai University, Nanjing, China. She has undertaken various projects on risk management of extreme flood and drought, water resource management system, water disaster emergency management, etc. She is the author of eight books and more than 200 articles. Her research interests include management science and system engineering, supply chain and optimal control, water resources system operations and management.

Jing Huang was born in Jiangsu, China, in 1986. She received the B.S. degree in information management and information system and the Ph.D. degree in management science and engineering from the Hohai University, Nanjing, China, in 2009 and 2015, respectively.

She is currently an Associate Professor with the Department of Management Science and Information Management, School of Business, Hohai University, Nanjing, China. Her research interests include risk assessment and management of flood and drought disaster, water resources management.

Lu Zhuo (Member, IEEE) received the M.Eng. and Ph.D. degrees in civil engineering from the University of Bristol, Bristol, U.K., in 2011 and 2016, respectively.

She is currently a Lecturer with the Department of Civil and Structural Engineering, University of Sheffield, Sheffield, U.K. She has authored and co-authored more than 30 peer-reviewed journal and conference papers including in high impact journals, such as *Journal of Hydrology*, *Hydrological Processes*, and *Hydrology and Earth System Sciences*. Her research interests include multiple natural hazards modeling and monitoring (e.g., floods, landslides, and earthquake), remote sensing of environment, and disaster risk management.