Stochastic Environmental Research and Risk Assessment (2021) 35:781–795
https://doi.org/10.1007/s00477-020-01916-8

ORIGINAL PAPER

Stream gauge network grouping analysis using community detection

Hongjun Joo[1] · Myungjin Lee[2] · Jongsung Kim[2] · Jaewon Jung[3] · Jaewon Kwak[4] · Hung Soo Kim[2]

Accepted: 22 October 2020 / Published online: 5 November 2020

Abstract

Stream gauging stations are important in hydrology and water science for obtaining water-related information, such as stage and discharge. However, for efficient operation and management, a more accurate grouping method is needed, which should be based on the interrelationships between stream gauging stations. This study presents a grouping method that employs community detection based on complex networks. The proposed grouping method was compared with the cluster analysis approach, which is based on statistics, to verify its adaptability. To achieve this goal, 39 stream gauging stations in the Yeongsan River basin of South Korea were investigated. The numbers of groups (clusters) in the study were two, four, six, and eight, which were determined to be suitable by fusion coefficient analysis. Ward's method was employed for cluster analysis, and multilevel modularity optimization was applied for community detection. A higher level of cohesion between stream gauging stations was observed in the community detection method at the basin scale and the stream link scale within the basin than in the cluster analysis. This suggests that community detection is more effective than cluster analysis in terms of hydrologic similarity, persistence, and connectivity. As such, these findings could be applied to grouping methods for efficient operation and maintenance of stream gauging stations.

Keywords Stream gauging station · Cluster analysis · Complex network · Community detection

1 Introduction

Stages or water levels are widely used in various fields, such as hydrology, water resource management, and environmental science, and one of the primary hydraulic structures for stage measurement is a stream gauging station (Sauer and Turnipseed 2010). Stage and flow data observed at stream gauging stations provide important water information for flood forecast warnings, operation of multipurpose dams, identification of available water resources, and operation of agricultural reservoirs. Consistent efforts have been made to achieve accurate stage measurements and quality control as high-quality stage data affect the reliability of flood, drought, water quality, and ecological management (Joo et al. 2019a, b), which require efficient management of stream gauging stations.

The operation and management of stream gauging stations require basin-scale analysis based on up-and-down streams, mainstreams, and tributaries. Specifically, stream gauge networks with small or mid-sized groups of gauging stations that share common characteristics within a basin are more favorable for the operation and management of the stations. More efficient management of stage data could be possible with accurate grouping methods based on the characteristics of stream gauging stations. In other words, an operation and maintenance strategy tailored to each group of stream gauging stations would allow for the management of stage data in problem situations. This requires a reasonable comparison and review of the grouping methods for stream gauging stations within a basin.

Popular grouping methods in hydrology include the well-known cluster analysis technique. Cluster analysis is

✉ Hung Soo Kim
sookim@inha.ac.kr

[1] Department of Construction Certification Center, Korea Institute of Civil Engineering and Building Technology, Ilsan 10223, Korea

[2] Department of Civil Engineering, Inha University, Incheon 22201, Korea

[3] Nakdong River Flood Control Office, Ministry of Environment, Busan 49300, Korea

[4] Center for Hydrology and Ecology, Incheon 22201, Korea

based on statistics and can identify differences between groups by bringing similar objects together and organizing them into a group. Cluster analysis based on time series data has been widely used in the field of hydrology (Kumar et al. 2015; Lin and Chen 2005; Kyung et al. 2007; Ouyang et al. 2010; Corduas 2011). Kumar et al. (2015) performed a cluster analysis to distinguish between seasonal periods accurately using a metric function based on the error distribution of seasonal data. Lin and Chen (2005) developed a time series prediction model for groundwater based on the self-organizing map (SOM), which is a two-dimensional map that directly identifies the number of clusters hidden in the radial basis function network (RBFN).

Kyung et al. (2007) used cluster analysis to create a Korean version of the hydrological drought severity-area-duration (SAD) curve, and high levels of severity were observed in the north and central areas along the eastern coast of the Korean Peninsula. Ouyang et al. (2010) performed a *K*-means cluster analysis on the mean monthly discharge, monthly maximum discharge, monthly amplitude, and monthly standard deviation from 1961 through 2000 for the Shaligunlanke Station in the Tarim River basin of China. The results showed that the annual process of daily discharge could be classified into five segments. Corduas (2011) performed a cluster analysis based on the bond energy algorithm (BEA), which is applicable to complex data arrays. The analysis was based on 89 hydrological time series data of mean daily discharge from rivers in Oregon and Washington in the United States.

Cluster analysis has also been conducted for catchment and hydrologic similarity classification, prediction of ungauged hydrological data, flood frequency analysis, hydrological modeling, and flood forecasting (Archfield et al. 2014; Auerbach et al. 2016; Boscarello et al. 2015; Isik and Singh 2008; Iyigun et al. 2013; Jingyi and Hall 2004; Kahya et al. 2008; Kileshye et al. 2012; Kuentz et al. 2017; Latt et al. 2015; Ouarda et al. 2008; Rao and Srinivas 2006; Rhee et al. 2008; Tercek et al. 2012; Unal et al. 2003). These studies were conducted using hierarchical clustering to combine similar clusters until eventually form a single group.

Network theory, which was invented in the eighteenth century (Euler 1741), was evolved to a next stage with complex network studies such as small-world networks, scale-free networks, network motifs, and community structure during the last two decades. It is one of the important tools that the actual application of complex network theory can describe a complicated and varied phenomenon (Sivakumar and Woldemeskel 2015). Complex network theory has also been applied recently in the field of hydrology (Rinaldo et al. 2006; Malik et al. 2012; Boers et al. 2013; Scarsoglio et al. 2013; Halverson and Fleming 2015; Sivakumar and Woldemeskel 2014; Fang et al. 2017; Han et al. 2018; Alarcòn and Lozano 2019; Kim et al. 2019). Community detection is a method based on network theory for grouping nodes that share similar or common goals. Fang et al. (2017) applied community detection in hydrology and organized communities using six methods (edge between centrality, greed algorithm, multilevel modularity optimization, leading eigenvector method, label propagation method, and the Walktrap method). The analysis was based on the similarity of daily streamflow for 1663 gauging stations across the Mississippi River in the United States. Halverson and Fleming (2015) organized communities of stream gauging stations located in the Coast Mountains of British Columbia and the Yukon in Canada according to seasonal flow regimes for each region and geographical proximity. Alarcòn and Lozano (2019) used Interbasin Transfer (IBT) for Spanish river basins to build a community structure consisting of seven small groups of two or three nodes.

Grouping methods have been used in the field of hydrology, including existing stream gauging stations, to identify differences in hydrological properties between groups. However, there have been insufficient efforts to review the accuracy and reliability of such grouping methods. Moreover, grouping methods have been recognized as a secondary process performed before the primary analysis. Comprehensive maintenance should be ensured across the board for stream gauging stations in the same group by applying more accurate grouping methods. Hydrologic aspects of gauging stations are directly affected by upstream gauging stations, and no individual station is independent.

The aim of this study is to present a grouping method using community detection based on complex networks. The proposed grouping method was compared with a statistical cluster analysis approach to verify its adaptability. This paper is organized as follows. Section 2 describes community detection and clustering methods. Multilevel modularity optimization and Ward's method that are used in this study and the methodology to select optimal number of group were also described. Section 3 constructs the stream gauge network that is consisting of the Node and Link using 39 stream gauging stations in the Youngsan River basin in South Korea and analyzes its community detection characteristics. Hierarchical cluster analysis was conducted according to the similarity between water levels. The grouping result and its characteristics were also analyzed. Based on the basin hydrology, the applicability of the complex network-based community detection method is compared with the statistical-based cluster analysis method. Finally, Sect. 4 provides a summary of the study.

2 Methods for community detection and cluster analysis

2.1 Network theory and community detection

2.1.1 Basic network theory

A network or graph is a set of points that are connected by a series of lines, as shown in Fig. 1. Points are called vertices or nodes, and lines are called edges or links. A network can be expressed as $G = \{P, E\}$, where P is a set of N nodes (P_1, P_2, …, P_N), and E is a set of n links. Figure 1 shows a network consisting of $N = 7$ nodes and $n = 8$ links. This network has a set of nodes $P = \{1, 2, 3, 4, 5, 6, 7\}$ and a set of links $E = \{(1,7), (2,7), (3,5), (3,7), (4,7), (5,6), (6,7)\}$.

Figure 1 shows the simplest form of a network that may appear in a more complex form. Examples of more complex networks include: (1) a network with one or more different types of nodes and links, (2) a network with different weights for different nodes and links, depending on the nodes and connection strength, (3) a network with cyclic or acyclic links, (4) a network with multi-links, self-links, and hyper-links, and (5) a network with two nodes that are separated from different types and operated independently in a separate type. Sivakumar (2015) provide a more detailed description of such networks.

Network characteristics can be studied in different ways. The key concepts of a network in the context of the modern theory of complex networks include centrality analysis, the clustering coefficient, degree distribution, and community detection. This study uses community detection for grouping the stream gauging stations within a basin.

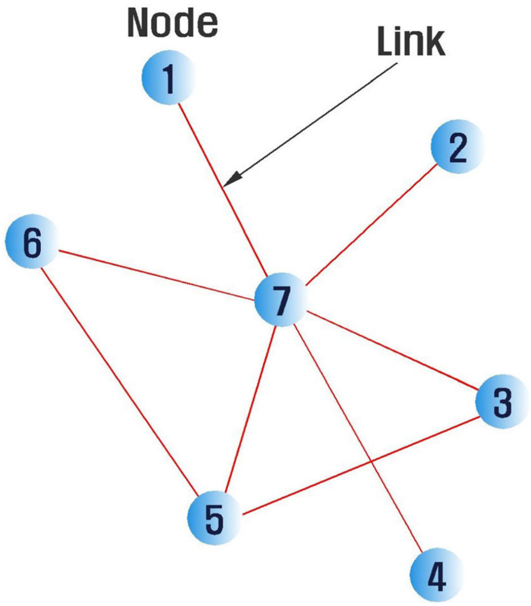


Fig. 1 Diagram of a network consisting of $N = 7$ nodes and $n = 8$ links

2.1.2 Community detection

In complex networks, nodes cluster together and form a group. The nodes in a group are closely connected, and the attributes of a group are typically independent of other groups. These groups are called communities, and finding communities is called community detection.

A community can also be called a cluster in a broad sense. However, the difference between clustering and community detection is that groups are formed only by the similarity of data in cluster analysis, while groups are formed by data similarity and by network theory and structure in the community detection method. Modularity is first used when constructing communities. The use of modularity allows us to quantify the differences between the number of connections in a community and the number of random connections, assuming a community within the entire network. The modularity equation proposed by Newman (2004) is commonly cited:

$$Q = \frac{1}{2m}\sum_{i,j}^{n}\left(a_{ij} - \frac{k_i k_j}{2m}\right)\delta\left(c_i, c_j\right) \tag{1}$$

where m is the total number of links, n is the total number of nodes, $\mathrm{a_{ij}}$ is the connectivity between nodes i and j, and k_i is the number of all connections to the nodes i and j. In addition, $\delta\left(\mathrm{c_i}, \mathrm{c_j}\right)$ is one when c_i and c_j are in the same community and zero when they are in different communities.

Modularity requires time-consuming calculations. Possible methods for overcoming this disadvantage include edge between centrality (Newman and Girvan 2004), the greedy algorithm (Clauset et al. 2004), the Walktrap method (Pons and Latapy 2005), the leading eigenvector method (Newman 2006), the label propagation method (Raghavan et al. 2007), and multilevel modularity optimization (Blondel et al. 2008). The aim of all methods is to improve modularity to optimize community detection.

The multilevel modularity optimization (or Louvain method) is the most recently developed modularity optimization and was employed for this study because it is designed to address the mentioned problems. For example, the greedy algorithm method and the multilevel modularity optimization method have the fastest community detection but poor optimization as they tend to create super communities. Multilevel modularity optimization consists of two phases that are repeated iteratively, which can be expressed as (Blondel et al. 2008):

$$\Delta Q = \left[\frac{\sum_{in} + k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \quad (2)$$

where $\sum_{in}$ is the sum of the weights of the links inside the community, $\sum_{tot}$ is the sum of the weights of the links incident to nodes in community, k_i is the sum of the weights of the links incident to node i, $k_{i,in}$ is the sum of the weights of the links from i to nodes in the community, and m is the sum of the weights of all the links in the network. Phase 1 forms communities using the improved modularity, and Phase 2 combines the communities created in Phase 1 into a block that is treated as a node. Next, the algorithm in Phase 1 again merges the newly modified networks. The model stops when no further changes occur in Phase 1 following Phase 2.

2.2 Cluster analysis

A cluster is based on the similar properties present in the interconnection between nodes. The task of classifying clusters based on similarity is called cluster analysis or clustering. Cluster analysis is based on statistics and can be classified into two types: hierarchical (agglomerative) clustering and partitional (divisive) clustering. Hierarchical clustering yields different cluster results step by step without predetermining the number of clusters. Partitional clustering is a method of specifying the number of clusters in advance. Based on these methods, the general procedure for network clustering is shown in Fig. 2.

Cluster analysis was employed to split stream gauging stations into groups based on stage data obtained from the stations. Hierarchical cluster analysis was applied because it can derive clusters of stream gauging stations without a predetermined number of clusters (Kaufman and Rousseeuw 2005). Several methods can be used to form clusters, such as single linkage, complete linkage, average linkage, and Ward's method.

Ward's method was used in this study. Unlike other methods, it is less sensitive to noise and outliers in data. Ward's method is very efficient and is widely used in many fields of science (Yoo et al. 2011). Other methods, such as single linkage, complete linkage, and average linkage, establish a group based on the similarity of each group using euclidean squared distance (L^2). But, Ward's method measures similarity using error sum of squares (ESS) when the two groups are combined. In other words, it conducts grouping that intends to minimize the increase of the ESS. In the initial clustering, all nodes are clustered one by one, and it can be expressed as $\mathrm{ESS}^{\mathrm{i}} = 0$ for all i. The ESS increases as further clustering occurs, which can be written as:

$$ESS = \sum_{i=1}^{g} ESS^{i} = \sum_{i=1}^{g}\left[\sum_{j=1}^{n_i}\sum_{k=1}^{p}\left(X_{kj}^{i} - \overset{-\;l}{\mathrm{X}_{l_{kj}}}\right)^2\right] \quad (3)$$

where $\overline{\mathrm{X}_{\mathrm{kj}}}^{\mathrm{i}}$ is the mean cluster for X_{k} in the ith cluster.

3 Application and results

3.1 Study area

The Yeongsan River basin is located in Southwestern South Korea (N 34° 40′ 16″–35° 29′ 01″, E 126° 26′ 12″–127° 06′ 07″). The need for maintenance based on stages at a stream gauging station has long been recognized for the Yeongsan River basin. Yeongsan River is one of the four major rivers in South Korea. It has a basin area of 3455 km^2 and a river length of 129.5 km and accommodates 39 stream gauging stations. Of the 39 stream gauging stations, 14 are deployed in the Yeongsan River, which is the main stream, and 25 are in the tributaries. These stations are under the supervision of the Ministry of Environment (ME) at the national level and reflect the importance of stream gauging station management and stage data. Figure 3 shows the location and the corresponding number of stream gauging stations within the basin.

3.2 Stage data setting

The data collection period must be predetermined as the complex network configuration for cluster analysis, and community detection is based on time series stage data obtained at each stream gauging station. To ensure the reliability of the stage data, a large amount of data for an extended period is needed (e.g., 30 years, which is generally considered suitable for hydrological analysis).

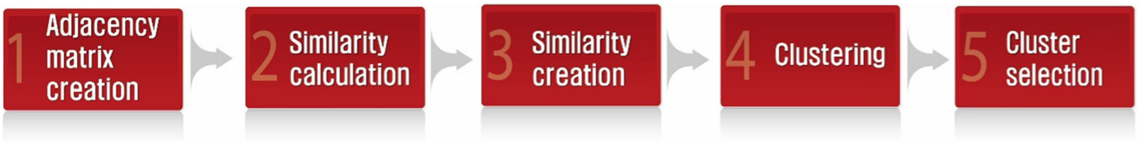


Fig. 2 General cluster analysis procedure

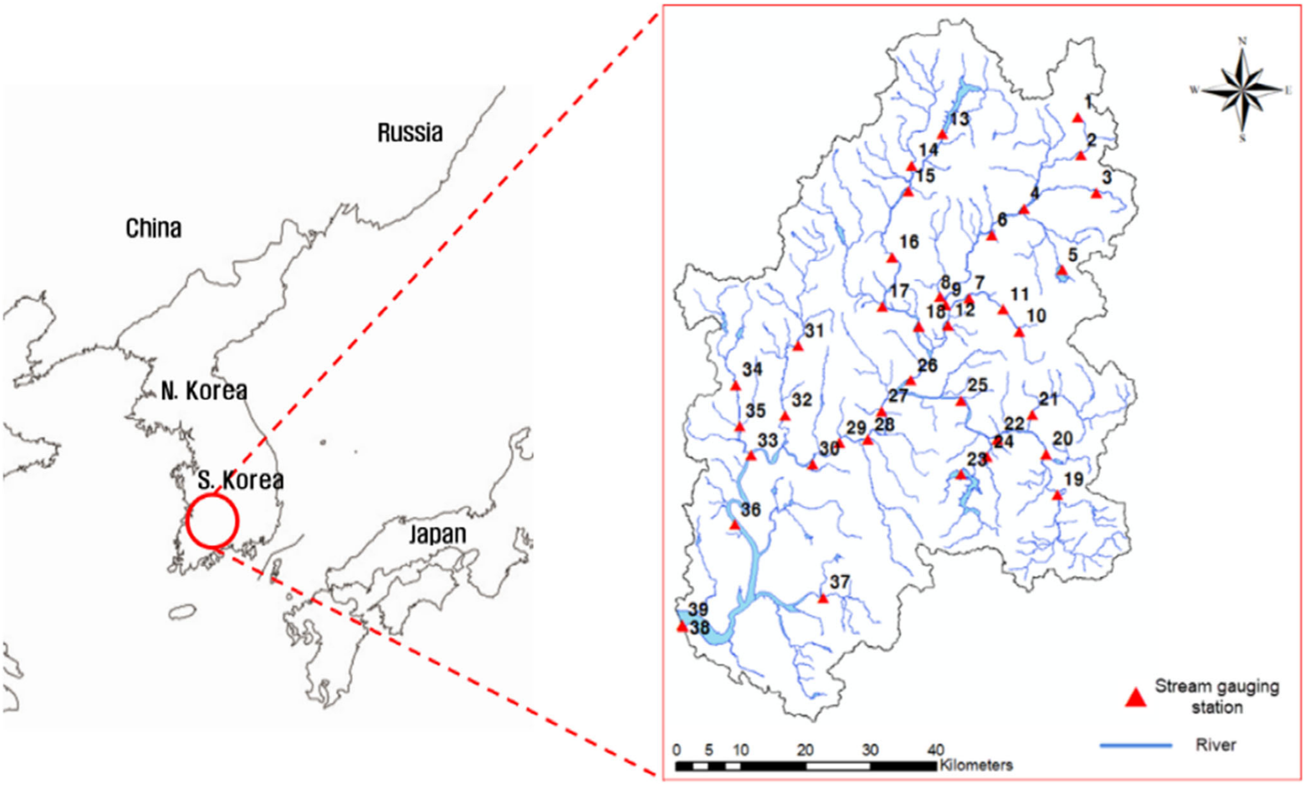


Fig. 3 Map of the Yeongsan River Basin and its stream gauging station locations (red triangles)

However, the required time period presents several challenges: (1) not all gauging stations have long period of data, (2) stage data of each gauging station are collected for different periods of time, (3) most gauging stations have missing data for one or more periods, and (4) data collected over a certain period of time at some gauging stations contain errors.

This study employs a grouping method based on the similarity of stage data, which requires an analysis of data collected over the same period of time from all stream gauging stations (Fang et al. 2017). Considering all the data, this study employed daily stage data collected over five years (January 2011 to December 2015), which provided consistent data for stream gauging stations. Figure 4 summarizes the distribution of stage data for each stream gauging station in a box plot.

Water levels remained constant at most stream gauging stations but showed high degrees of variability at some stations. Stream gauging stations exhibited higher variability in water levels over short and long periods of time, which can be attributed to the fact that the stations are located directly downstream of dams or near weirs (Stations 1, 5, 13, 17, 23, and 30), which are operated for flow control. Moreover, most outliers observed at each gauging station resulted from a rapid rise in the stage caused by localized heavy rainfall during the flood season. Water level events that vary significantly with regional, meteorological, and manmade factors were also included for analysis as these events have effects on other gauging stations at the watershed level.

3.3 Community detection of stream gauging stations based on network theory

In complex networks, stream gauging stations can be represented as nodes without links that connect them. Stream gauging stations are installed along a water system that only serves as a means of accommodating the stations, not as a link to connect them. Therefore, links should be constructed to connect each stream gauging station based on the correlation of stage data. This requires a focus on the network configuration that changes with threshold values (T) of similarity in stage data to investigate the diverse communities. Therefore, a complex network was constructed with threshold values based on the correlation between 0.1 and 1.0, and community formation was carried out for the 39 stream gauging stations using multilevel modularity optimization (Fig. 5). The results showed that networks configured differently depending on the threshold values also had different community formation.

The number of communities based on threshold value could be up to 39 that is the total number of stream gauging

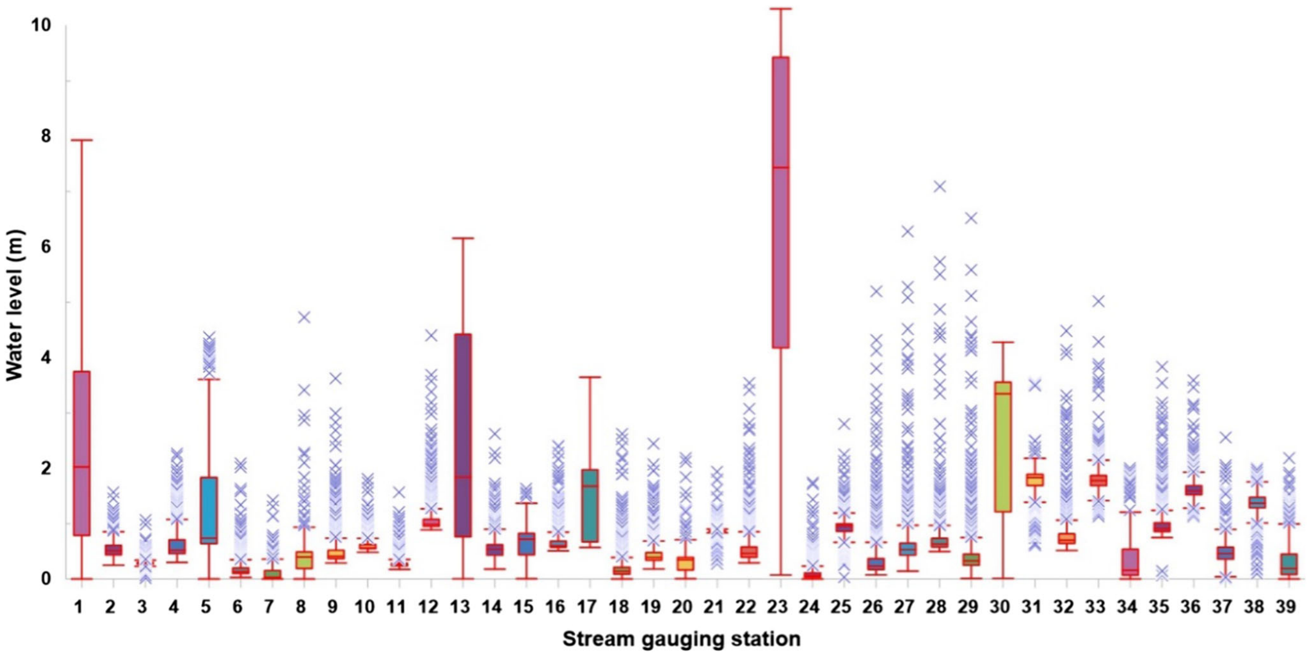


Fig. 4 Distribution of daily stage data obtained from 39 stream gauging stations (January 2011 to December 2015)

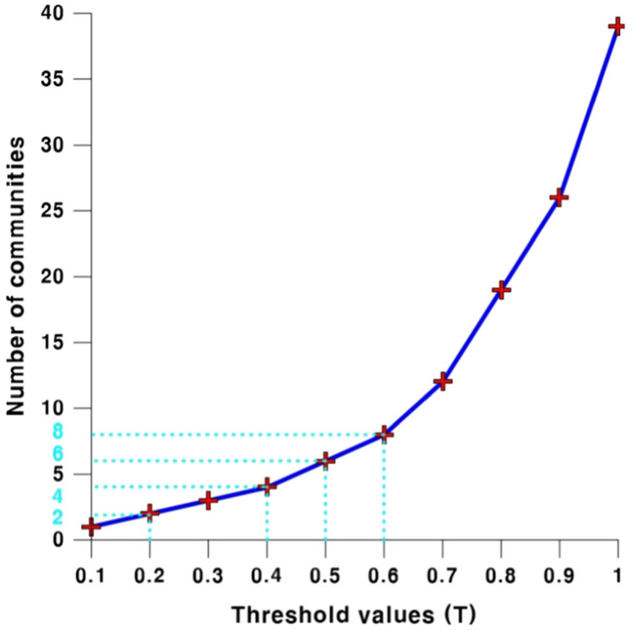


Fig. 5 Change in number of communities according to threshold values (T)

stations. The proper number of communities could be estimated when the modularity are maximized on the multilevel modularity optimization method. The appropriate number of communities could be estimated when the modularity is maximized on the multilevel modularity optimization method (see Eq. 2). The maximum modularity (Q = 0.518) is calculated when communities are eight, and this study employs 2, 4, 6, and 8 communities to consider each event of community number. It is equivalent to 0.2, 0.4, 0.5 and 0.6 in threshold values.

Groups of four events represented different types of links and exhibited the typical structure of complex networks. Figure 6 shows the community results for different stream gauging stations, where boxes of the same color indicate that they belong to the same community group. Figure 7 shows the community results for different locations of the stream gauging stations. The solid lines between stream gauging stations are based on threshold values and determine the network structure.

The results of community detection showed that clustering mostly took place in a group (the green group) across all group events, similar to the cluster analysis. However, the results were more centralized in community detection. Furthermore, stream gauging stations (nodes) that are not connected by links did not always organize into different communities. In addition to nodes, community formation in the network involves other factors, such as the number of links for neighboring adjacent nodes and the degree and intensity of connection between links. In other words, even if nodes are not linked to each other due to a lack of similarity in stage data, they can still be organized into the same community if indirectly connected by other nodes.

3.4 Cluster analysis of stream gauging station based on stage data

A hierarchical cluster analysis was performed for the 39 stream gauging stations. The resulting dendrogram is

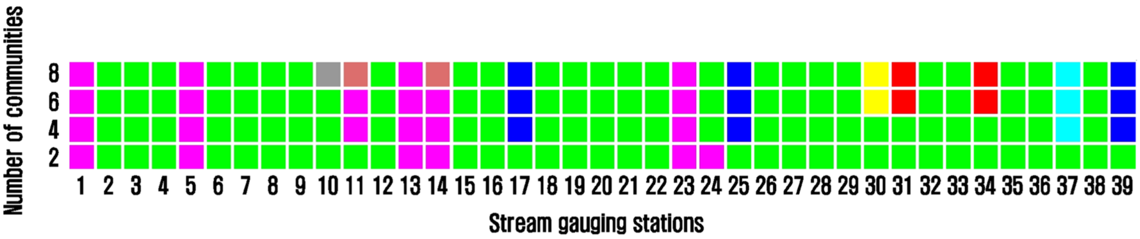
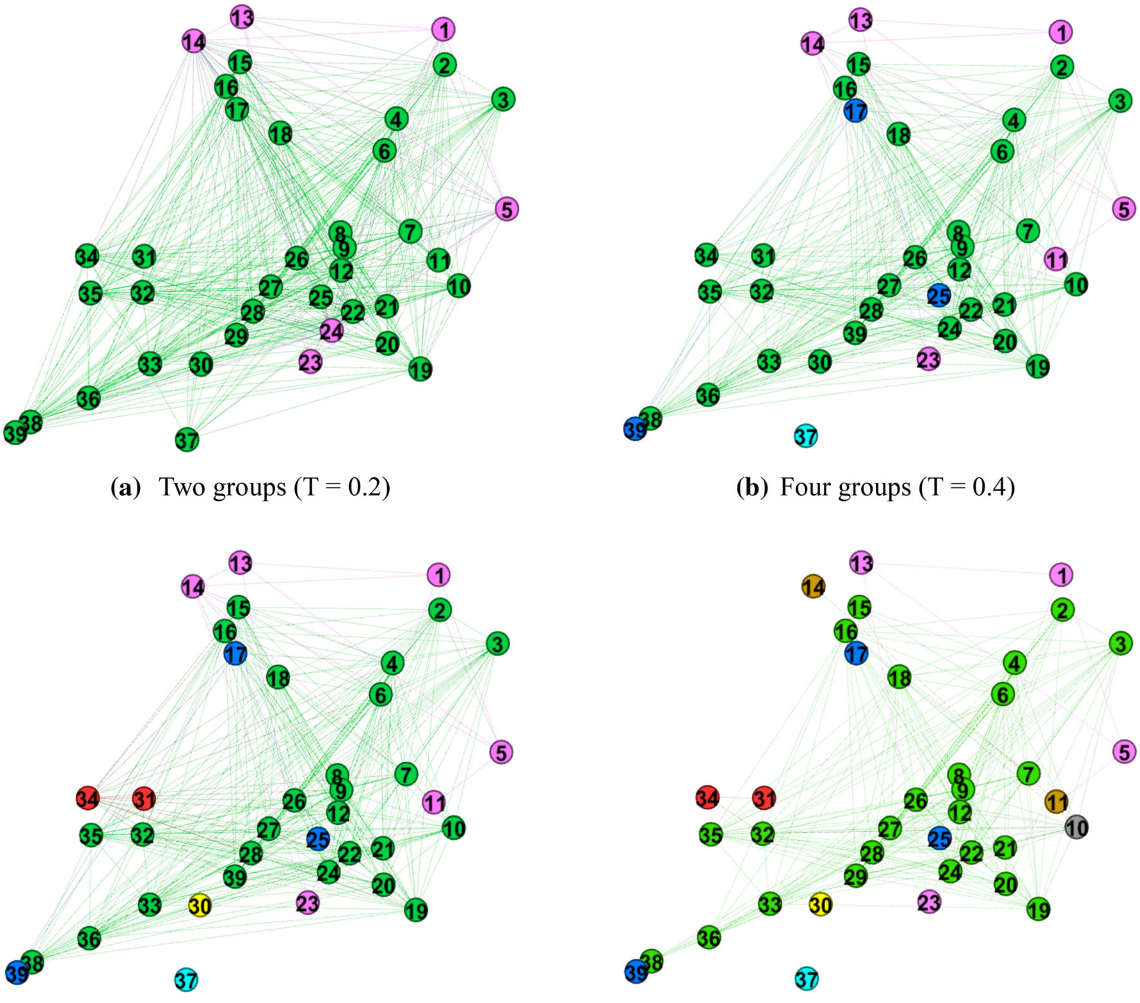


Fig. 6 Community detection for stream gauging stations (number of groups: two, four, six, and eight)



(a) Two groups (T = 0.2)

(b) Four groups (T = 0.4)

(c) Six groups (T = 0.5)

(d) Eight groups (T = 0.6)

Fig. 7 Diagram of community structure according to the relative location of stream gauging stations (number of groups: two, four, six, and eight)

shown in Fig. 8. The maximum number of clusters is 39, which is the total number of stream gauging stations investigated. The clustering can be divided into 10 steps depending on the number of groups.

Determining the number of clusters is extremely challenging in a typical cluster analysis. This also applies to the case where the number of clusters is determined using the similarity of stage data for efficient management of stream gauging stations. Many studies have been conducted to determine the appropriate number of clusters (Aaker et al. 2001). When using Ward's method, the ESS variance with the number of clusters is represented by the fusion coefficient (Eq. 3). The fusion coefficient derived at every stage of clustering was effectively used for this purpose. The fusion coefficient is estimated by considering the distances between clusters at each stage of clustering, so its value can be used to determine how newly made clusters differ from each other. That is, if the fusion coefficient shows a

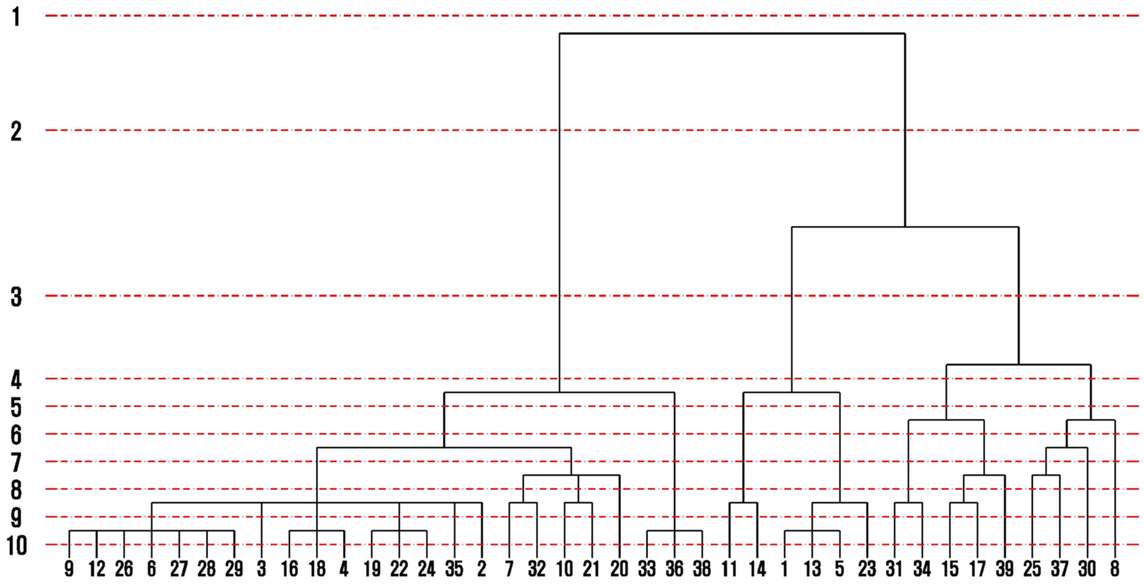


Fig. 8 Derivation of clustering dendrogram for stream gauging stations

significant increase when decreasing the number of clusters, two relatively different clusters can be made into a single cluster at the current stage (Aldenderfer and Blashfield 1984; Yoo et al. 2011).

Figure 9 summarizes the derived fusion coefficients. Like community detection, the results show that a significant change in the fusion coefficient is observed when the number of clusters is equal to eight, which suggests that the appropriate number of clusters will be eight or less. Therefore, the changes in four events were investigated, which consisted of two, four, six, and eight clusters (groups) for the same comparison with the result of community detection.

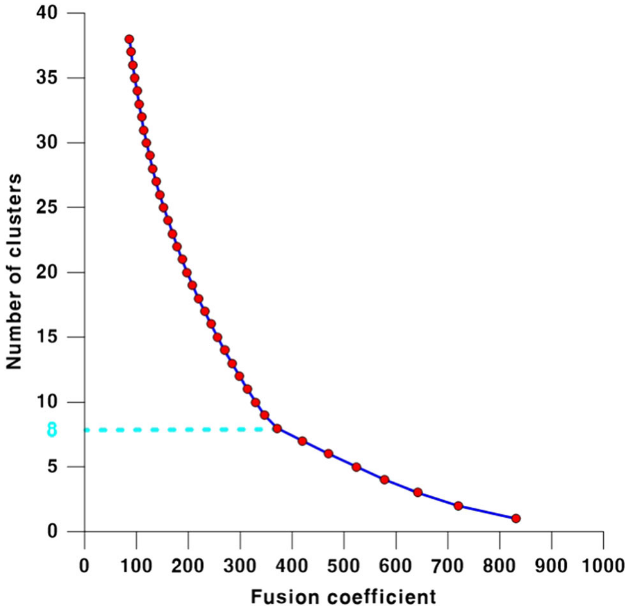


Fig. 9 Changes in the fusion coefficient according to the number of clusters

The results of cluster analysis performed on four events based on fusion coefficients are illustrated in Figs. 10 and 11. The boxes of the same color in Fig. 7 indicate that they belong to the same cluster group, while Fig. 8 shows the clustering results according to the relative location of the stream gauging stations.

The results show that clustering mostly took place in a group (the green group) across all group events. This can be interpreted as indicating possible similarity of stage data at the basin level. However, at the same time, it indicates that stream gauging stations in the target basin have somewhat unclear clustering without distinguishing characteristics compared community detection (comparing Figs. 8 and 11). For nearby stream gauging stations, water levels are often similar to each other in general. In contrast, stream gauging stations within the target river basin were often found to belong to different groups, despite their proximity. This can be attributed to different stage data resulting from topographic conditions, such as river bed elevation, despite the close proximity of stream gauging stations. A more quantitative analysis is required to provide a more detailed comparison between community detection and cluster analysis.

3.5 Comparison and discussion of grouping methods based on basin hydrology

The grouping methods for stream gauging stations based on cluster analysis and community detection were compared in terms of basin hydrology and evaluated for suitability.

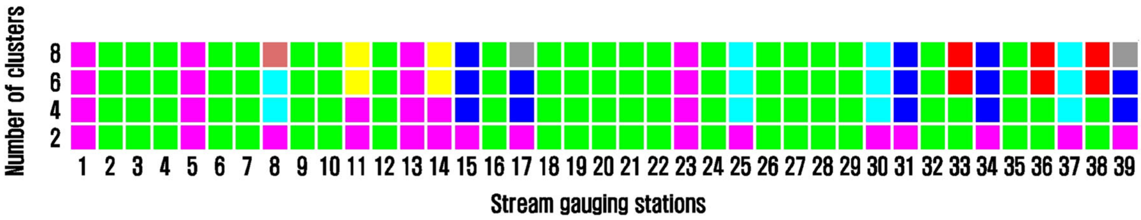


Fig. 10 Cluster analysis of stream gauging stations (number of groups: two, four, six, and eight)

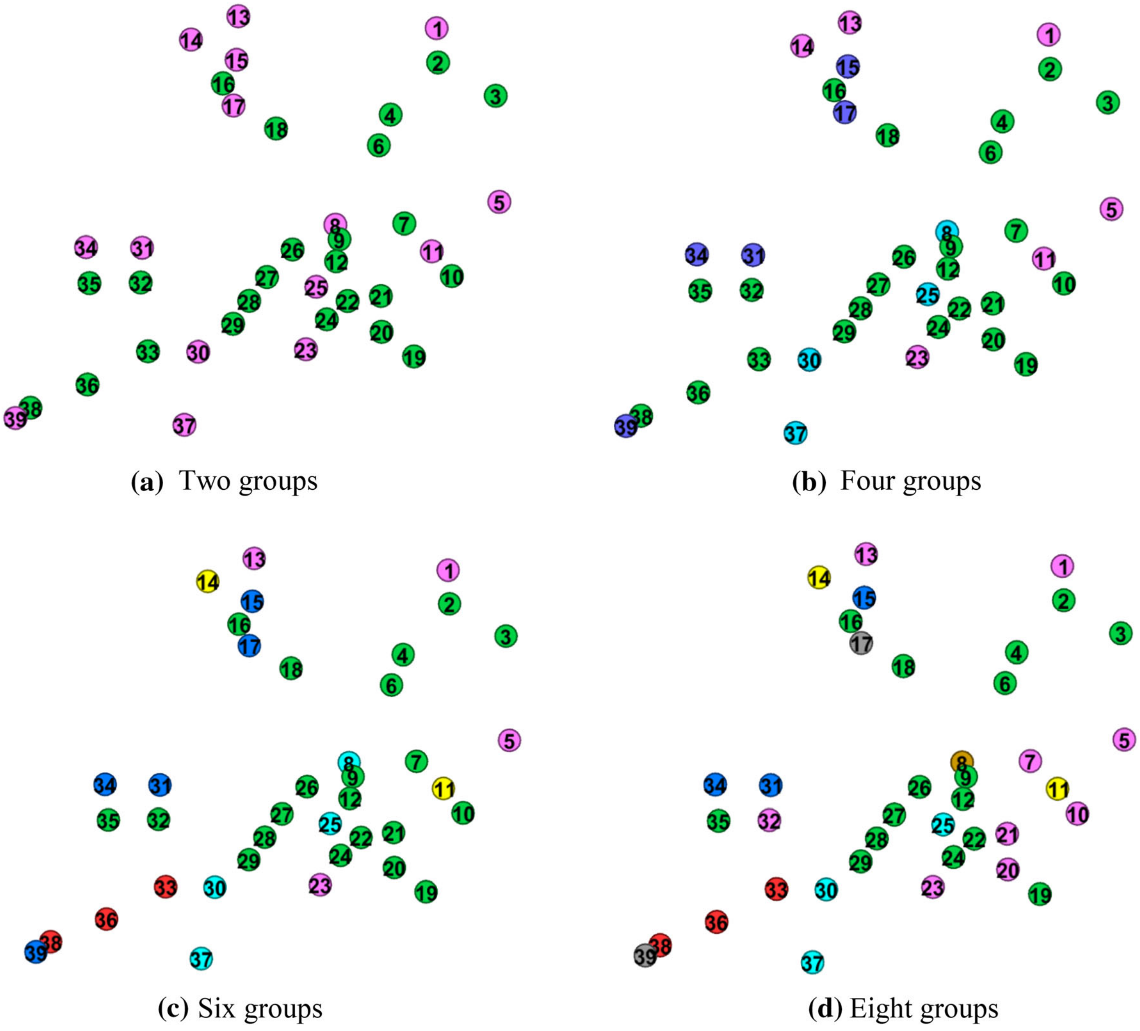


Fig. 11 Cluster analysis of stream gauging station (number of groups: two, four, six, and eight)

As shown in Table 1, when the number of groups is two, four, six, and eight, the number of stream gauging stations that can be included in communities and clusters is 156 in total (39 stations × 4 events). As mentioned in Sect. 3.4, gauging stations were found to belong to one group (the green group) in most cases and for both grouping methods. Stations were more likely to belong to one group for communities than clusters, and very few stations changed to another group as grouping took place. This indicates that community detection in the stream gauge networks at the basin level resulted in relatively high levels of similarity—that is, both direct and indirect cohesion occurred among stream gauging stations, in contrast to cluster analysis.

For hydrologic comparison by group and grouping method, it is necessary to investigate the groups connected by the same stream links between gauging stations. Therefore, the changes in stream gauging stations according to grouping method were studied for a total of 12 stream links. The main stream of the Yeongsan River within the basin was set as stream link Index A, and the

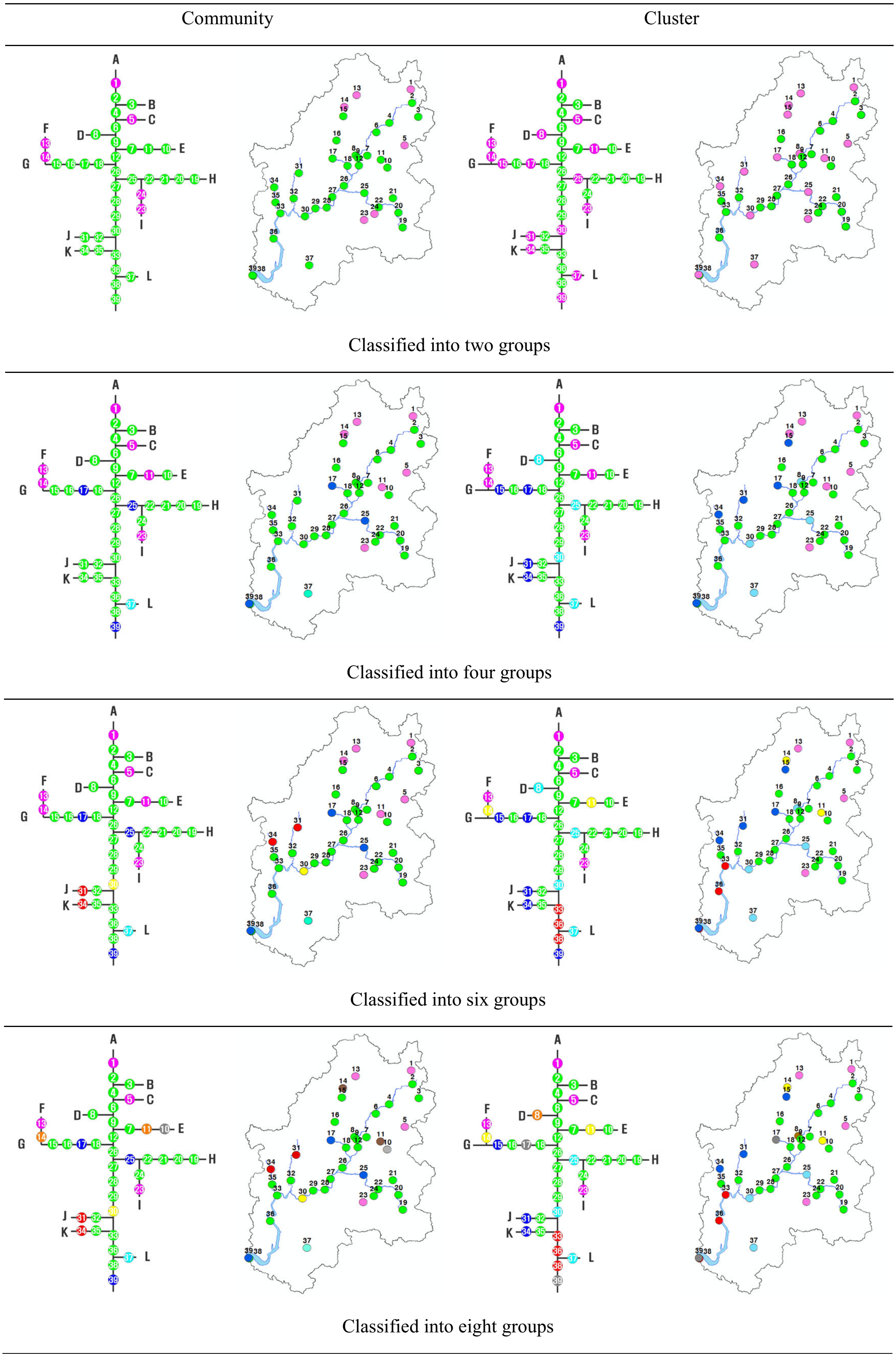
Table 1 Number of stream gauging stations selected by group and grouping method

No. of groups	Community								Cluster							
	🟩	🟪	🟦	🔷	🟥	🟨	⬛	🟧	🟩	🟪	🟦	🔷	🟥	🟨	⬛	🟧
2	33	6	–	–	–	–	–	–	24	15	–	–	–	–	–	–
4	29	6	3	1	–	–	–	–	24	6	5	4	–	–	–	–
6	26	6	3	1	2	1	–	–	21	4	5	4	3	2	–	–
8	25	4	3	1	2	1	1	2	21	4	3	3	3	2	2	1
Sub sum	113	22	9	3	4	2	1	2	91	29	13	10	6	4	2	1
Total sum	156								156							

Table 2 Grouping and gauging station structure for different stream links (Column ①: the number of gauging stations in the group with the most stations; Column ②: the number of other gauging stations)

Stream link index	No. of groups: two						No. of groups: four					
	Community			Cluster			Community			Cluster		
	Total	①	②	Total	①	②	Total	①	②	Total	①	②
A	15	14	1	15	12	3	15	13	2	15	12	3
B	1	1	0	1	1	0	1	1	0	1	1	0
C	1	1	0	1	1	0	1	1	0	1	1	0
D	1	1	0	1	1	0	1	1	0	1	1	0
E	3	3	0	3	2	1	3	2	1	3	2	1
F	2	2	0	2	2	0	2	2	0	2	2	0
G	4	4	0	4	2	2	4	3	1	4	2	2
H	5	5	0	5	4	1	5	4	1	5	4	1
I	2	2	0	2	1	1	2	1	1	2	1	1
J	2	2	0	2	1	1	2	2	0	2	1	1
K	2	2	0	2	1	1	2	2	0	2	1	1
L	1	1	0	1	1	0	1	1	0	1	1	0

Stream link index	No. of groups: six						No. of groups: eight					
	Community			Cluster			Community			Cluster		
	Total	①	②	Total	①	②	Total	①	②	Total	①	②
A	15	12	3	15	9	6	15	12	3	15	9	6
B	1	1	0	1	1	0	1	1	0	1	1	0
C	1	1	0	1	1	0	1	1	0	1	1	0
D	1	1	0	1	1	0	1	1	0	1	1	0
E	3	2	1	3	2	1	3	1	2	3	2	1
F	2	2	0	2	1	1	2	1	1	2	1	1
G	4	3	1	4	2	2	4	3	1	4	2	2
H	5	4	1	5	4	1	5	4	1	5	4	1
I	2	1	1	2	1	1	2	1	1	2	1	1
J	2	1	1	2	1	1	2	1	1	2	1	1
K	2	1	1	2	1	1	2	1	1	2	1	1
L	1	1	0	1	1	0	1	1	0	1	1	0



Community

Cluster

Classified into two groups

Classified into four groups

Classified into six groups

Classified into eight groups

◀**Fig. 12** Diagram and basin map based on cluster analysis and community detection

Table 3 Calculation and comparison of cohesion by group and method (italic cells mean higher cohesion in the same group)

Stream link index	Community (%)				Cluster (%)			
	2	4	6	8	2	4	6	8
A	*93.3*	*86.7*	*80*	*80*	80	80	60	60
E	*100*	66.7	66.7	33.3	66.7	66.7	66.7	*66.7*
F	100	100	100	50	100	100	50	50
G	*100*	*75*	*75*	*75*	50	50	50	50
H	*100*	80	80	80	80	80	80	80
I	*100*	50	50	50	50	50	50	50
J	*100*	*100*	50	50	50	50	50	50
K	*100*	*100*	50	50	50	50	50	50

remaining tributaries were set as stream links B through L (Table 2 and Fig. 12). Table 2 shows the stream gauging stations located in stream links A through L and the group structure based on cluster and community.

The total value in Table 2 represents the number of stream gauging stations located in each stream link. In addition, column (1) represents the number of gauging stations in the group that has the largest number of stations. This value can be used to analyze the cohesion between gauging stations for different stream links in comparison to the entire stations. Column (2) represents the remaining stations other than those listed in column (1). Therefore, the sum of the number of gauging stations in the group consisting of the most stations and the number of the remaining stations is the total number of gauging stations located in each stream link.

As shown in Table 2, the numbers of stream gauging stations located in stream links A through L are 15, 1, 1, 1, 3, 2, 4, 5, 2, 2, 2, and 1, respectively. This indicates that the majority of gauging stations in the Yeongsan River Basin is located in the main stream (stream link A) of the Yeongsan River. In the stream link A, the number of stream gauging stations (column (1)) in the group consisting of the most stations was 12, 12, 9, and 9 for groups two, four, six, and eight, respectively. These values are smaller than those of the community detection method (14, 13, 12, and 12). This indicates that the stream gauging stations in stream link A are densely connected by the stream link and exhibit relatively strong cohesion in the application of community detection.

In contrast, when the number of groups was eight, the number of stream gauging stations in the group consisting of the most stations in the E stream link was two for cluster analysis and one for community detection. This indicates that cluster analysis identifies gauging stations with strong group cohesion. As a result of community detection, three gauging stations were organized into different groups.

Figure 12 shows a diagram and basin map of gauging station grouping for different stream links. In both methods, stream gauging stations such as 1, 17, 25, and 39 often did not belong to the same group as the nearby stations, which can be attributed to the dissimilarity of the stage due to the topographic attributes rather than meteorological effects. Moreover, stream gauging station 8 located in stream link D could not be grouped into the main stream of the Yeongsan River (stream link A) in the cluster analysis due to lack of similarity in stage data between the two stream links.

In contrast, station 8 was organized into the same group in the community detection method based on network theory. This indicates that stream link D affects the main stream of the Yeongsan River despite the lack of similarity in stage data. This also suggests that station 8 must be operated and managed in connection with gauging stations located on the main stream in community detection.

To make a quantitative comparison between the two methods at the stream scale cohesion among the gauging stations was expressed as:

$$G_c = \left(\frac{S_M}{ST}\right) \times 100 \tag{4}$$

where G_C is the cohesion according to grouping methods, S_M is the number of gauging stations in the group consisting of the most stations in a certain stream link, and S_T is the total number of gauging stations in a certain stream link, which can be expressed as (column (1)/Total) × 100 based on Table 2. The cohesion calculated using this method is shown in Table 3; the stream links consisting of a single station (B, C, D, and L) were excluded because the stream-scale comparison was irrelevant in this case.

The results showed that community detection identified highly cohesive gauging stations in most stream links compared to cluster analysis. This appears more evident in the stream links containing many gauging stations, such as

A and G, throughout the entire group. In other words, gauging stations in the same stream link are more likely to be grouped together in the community detection method. Other groups (groups three, five, seven) that were not considered in this study show the same results. This indicates that community detection based on the network structure of nodes and links is more suitable than cluster analysis in hydrology.

The streamflow in a stream link generally exhibits a high degree of hydrologic similarity, persistence, and connectivity, indicating that gauging stations are not independent but are closely related to each other. Therefore, a network-based community detection method that deals with communities with high cohesion would be a better alternative to the cluster analysis method, which is simply based on data correlation. In the community detection method, stream gauging stations that are not organized into major groups at the stream link scale require special maintenance tailored to the attributes of the non-major groups. The results of this study are expected to serve as an appropriate selection method for a small number of stream gauging stations with different characteristics.

4 Conclusions

This study evaluated the adaptability of community detection based on complex networks as a grouping method for efficient operation and maintenance of stream gauging stations. To achieve this goal, 39 stream gauging stations in the Yeongsan River Basin of South Korea were investigated using the community detection method. These results were compared with statistical cluster analysis results. For community detection and cluster analysis, multilevel modularity optimization and Ward’s method were employed. The number of groups was set to two, four, six, and eight based on modularity and fusion coefficient analysis, respectively.

The results showed that communities are more likely to be arranged into a group in the community detection method than in the cluster analysis. This indicates that the grouping of stream gauging stations at the basin scale has higher levels of cohesion in community detection than in cluster analysis. For comparison purposes in terms of hydrological conditions, the changes of the stream gauging stations located in a total of 12 stream links (A through L) and including the main stream of the Yeongsan River were investigated for different groups and methods. Higher levels of cohesion among the gauging stations were observed in the community detection method in most stream links.

High cohesion in a stream link means a high degree of hydrologic similarity, persistence, and connectivity. In turn, this makes the community detection method a better candidate for grouping as it can successfully simulate the general stream attributes. The present findings are expected to serve as a grouping method for the comprehensive management of stream gauging stations. This study analyzed only for water level. However, we may need further work for the integrated grouping of water level and other hydrologic components such as water utilization, flow control, environment, and gauging station impact.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korean government (MSIT) (No. 2017R1A2B3005695).

References

Aaker DA, Kumar V, Day G (2001) Marketing research. Wiley, New York

Alarcòn RR, Lozano S (2019) A complex network analysis of Spanish river basins. J Hydrol 578:124065. https://doi.org/10.1016/j.jhydrol.2019.124065

Aldenderfer MS, Blashfield RK (1984) Cluster analysis, series of quantitative applications in the social sciences, vol 44. Sage Univ., Beverly Hills, pp 38–43. https://doi.org/10.4135/9781412983648

Archfield SA, Kennen JG, Carlisle DM, Wolock DM (2014) An objective and parsimonious approach for classifying natural flow regimes at a continental scale. River Res Appl 30(9):1166–1183. https://doi.org/10.1002/rra.2710

Auerbach DA, Buchanan BP, Alexiades AV, Anderson EP, Encalada AC, Larson EI, McManamay RA, Poe GL, Walter MT, Flecker AS (2016) Towards catchment classification in data-scarce regions. Ecohydrology. https://doi.org/10.1002/eco.1721

Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech 10:P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Boers N, Bookhagen B, Marwan N, Kurths J, Marengo J (2013) Complex networks identify spatial patterns of extreme rainfall events of the South American Monsoon System. Geophys Res Lett 40(16):4386–4392. https://doi.org/10.1002/grl.50681

Boscarello L, Ravazzani G, Cislaghi A, Mancini M (2015) Regionalization of flow-duration curves through catchment classification with streamflow signatures and physiographic–climate indices. J Hydrol Eng 21(3):05015027. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001307

Clauset A, Newman MEJ, Moore C (2004) Finding community structure in very large networks. Phys Rev E 70:066111. https://doi.org/10.1103/PhysRevE.70.066111

Corduas M (2011) Clustering streamflow time series for regional classification. J Hydrol 407(1):73–80. https://doi.org/10.1016/j.jhydrol.2011.07.008

Euler L (1741) Solutio problematis ad geometriam situs pertinentis. Comment Acad Sci Petropolitanae 8:128–140

Fang K, Sivakumar B, Woldemeskel FM (2017) Complex networks, community structure, and catchment classification in a large-scale river basin. J Hydrol 545:478–493. https://doi.org/10.1016/j.jhydrol.2016.11.056

Halverson MJ, Fleming SW (2015) Complex network theory, streamflow, and hydrometric monitoring system design. Hydrol Earth Syst Sci 19:3301–3318. https://doi.org/10.5194/hess-19-3301-2015

Han X, Sivakumar B, Woldemeskel FM, Aguilar MG (2018) Temporal dynamics of streamflow: application of complex networks. Geosci Lett 5:10. https://doi.org/10.1186/s40562-018-0109-8

Isik S, Singh VP (2008) Hydrologic regionalization of watersheds in Turkey. J Hydrol Eng 13(9):824–834. https://doi.org/10.1061/(ASCE)1084-0699(2008)13:9(824)

Iyigun C, Türkeş M, Batmaz İ, Yozgatligil C, Purutçuoğlu V, Koç EK, Öztürk MZ (2013) Clustering current climate regions of Turkey by using a multivariate statistical method. Theor Appl Climatol 114(1–2):95–106. https://doi.org/10.1007/s00704-012-0823-7

Jingyi Z, Hall MJ (2004) Regional flood frequency analysis for the Gan-Ming River basin in China. J Hydrol 296(1):98–117. https://doi.org/10.1016/j.jhydrol.2004.03.018

Joo HJ, Jun HD, Lee JH, Kim HS (2019) Assessment of a stream gauge network using upstream and downstream runoff characteristics and entropy. Entropy. https://doi.org/10.3390/e21070673

Joo HJ, Lee JH, Jun HD, Kim KT, Hong SJ, Kim JW, Kim HS (2019) Optimal stream gauge network design using entropy theory and importance of stream gauge stations. Entropy. https://doi.org/10.3390/e21100991

Kahya E, Demirel MC, Beg O (2008) Hydrologic homogeneous regions using monthly streamflow in Turkey. Earth Sci Res J 12(2):181–193

Kaufman L, Rousseeuw PJ (2005) Finding groups in data: an introduction to cluster analysis. Wiley, New York

Kileshye Onema JM, Taigbenu AE, Ndiritu J (2012) Classification and flow prediction in a data-scarce watershed of the equatorial Nile region. Hydrol Earth Syst Sci 16(5):1435. https://doi.org/10.5194/hess-16-1435-2012

Kim KH, Joo HJ, Han DG, Kim SJ, Lee TW, Kim HS (2019) On complex network construction of rain gauge stations considering nonlinearity of observed daily rainfall data. Water. https://doi.org/10.3390/w11081578

Kuentz A, Arheimer B, Hundecha Y, Wagener T (2017) Understanding hydrologic variability across Europe through catchment classification. Hydrol Earth Syst Sci 21:2863–2879. https://doi.org/10.5194/hess-21-2863-2017

Kumar R, Goel NK, Chatterjee C, Nayak PC (2015) Regional flood frequency analysis using soft computing techniques. Water Res Manag 29(6):1965–1978. https://doi.org/10.1007/s11269-015-0922-1

Kyung MS, Kim SD, Kim BK, Kim HS (2007) Construction of hydrological drought severity-area-duration curves using cluster analysis. J Korean Soc Civ Eng 27(3B):267–276

Latt ZZ, Wittenberg H, Urban B (2015) Clustering hydrological homogeneous regions and neural network-based index flood estimation for ungauged catchments: an example of the Chindwin River in Myanmar. Water Res Manag 29(3):913–928. https://doi.org/10.1007/s11269-014-0851-4

Lin GF, Chen LH (2005) Time series forecasting by combining the radial basis function network and the self-organizing map. Hydrol Process 19(10):1925–1937. https://doi.org/10.1002/hyp.5637

Malik N, Bookhagen B, Marwan N, Kurths J (2012) Analysis of spatial and temporal extreme monsoonal rainfall over South Asia using complex networks. Clim Dyn 39(3–4):971–987. https://doi.org/10.1007/s00382-011-1156-4

Newman MEJ (2004) Fast algorithm for detecting community structure in networks. Phys Rev E 69:066133. https://doi.org/10.1103/PhysRevE.69.066133

Newman MEJ (2006) Finding community structure using the eigenvectors of matrices. Phys Rev E 74:036104. https://doi.org/10.1103/PhysRevE.74.036104

Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69:026113. https://doi.org/10.1103/PhysRevE.69.026113

Ouarda TBMJ, Bâ KM, Diaz-Delgado C, Cârsteanu A, Chokmani K, Gingras H, Quentin E, Trujillo E, Bobée B (2008) Intercomparison of regional flood frequency estimation methods at ungauged sites for a Mexican case study. J Hydrol 348(1):40–58. https://doi.org/10.1016/j.jhydrol.2007.09.031

Ouyang R, Ren L, Cheng W, Zhou C (2010) Similarity search and pattern discovery in hydrological time series data mining. Hydrol Process 24(9):1198–1210. https://doi.org/10.1002/hyp.7583

Pons P, Latapy M (2005) Computing communities in large networks using random walks. Lect Notes Comput Sci 3733:284–293. https://doi.org/10.1007/11569596_31

Raghavan UN, Albert R, Kumara S (2007) Near linear time algorithm to detect community structures in large-scale networks. Phys Rev E 76:036106. https://doi.org/10.1103/PhysRevE.76.036106

Rao AR, Srinivas VV (2006) Regionalization of watersheds by hybrid-cluster analysis. J Hydrol 318(1):37–56. https://doi.org/10.1016/j.jhydrol.2005.06.003

Rhee J, Im J, Carbone GJ, Jensen JR (2008) Delineation of climate regions using in-situ and remotely-sensed data for the Carolinas. Remote Sens Environ 112(6):3099–3111. https://doi.org/10.1016/j.rse.2008.03.001

Rinaldo A, Banavar JR, Maritan A (2006) Trees, networks, and hydrology. Water Resour Res 42:W06D07. https://doi.org/10.1029/2005WR004108

Sauer VB, Turnipseed DP (2010) Stage measurement at gauging stations. In: U.S. geological survey techniques and methods, Chapter 7 of Book 3, Section A. U.S. Geological Survey, USA, p 60. https://doi.org/10.3133/tm3A7

Scarsoglio S, Laio F, Ridolfi L (2013) Climate dynamics: a network-based approach for the analysis of global precipitation. PLoS ONE 8(8):e71129. https://doi.org/10.1371/journal.pone.0071129

Sivakumar B (2015) Networks: a generic theory for hydrology? Stoch Environ Res Risk Assess 29:761–771. https://doi.org/10.1007/s00477-014-0902-7

Sivakumar B, Woldemeskel FM (2015) A network-based analysis of spatial rainfall connections. Environ Model Softw 69:55–62. https://doi.org/10.1016/j.envsoft.2015.02.020

Sivakumar B, Woldemeskel FM (2014) Complex networks for streamflow dynamics. Hydrol Earth Syst Sci 18:4565–4578. https://doi.org/10.5194/hess-18-4565-2014

Tercek MT, Gray ST, Nicholson CM (2012) Climate zone delineation: evaluating approaches for use in natural resource management. Environ Manag 49(5):1076–1091. https://doi.org/10.1007/s00267-012-9827-4

Unal Y, Kindap T, Karaca M (2003) Redefining the climate zones of Turkey using cluster analysis. Int J Climatol 23(9):1045–1055. https://doi.org/10.1002/joc.910

Yoo CS, Ku HJ, Kim KW (2011) Use of a distance measure for the comparison of unit hydrographs: application to the stream gauge network optimization. J Hydrol Eng 16:880–890. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000393

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.