Hydrol. Earth Syst. Sci. Discuss., 11, 7255–7289, 2014
www.hydrol-earth-syst-sci-discuss.net/11/7255/2014/
doi:10.5194/hessd-11-7255-2014

Hydrology and Earth System Sciences
Discussions
Open Access

Complex networks for streamflow dynamics

B. Sivakumar[1,2] **and F. M. Woldemeskel**[1]

[1]School of Civil and Environmental Engineering, The University of New South Wales, Sydney, Australia
[2]Department of Land, Air and Water Resources, University of California, Davis, USA

Received: 3 June 2014 – Accepted: 15 June 2014 – Published: 2 July 2014

Correspondence to: B. Sivakumar (s.bellie@unsw.edu.au)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

Streamflow modeling is an enormously challenging problem, due to the complex and nonlinear interactions between climate inputs and landscape characteristics over a wide range of spatial and temporal scales. A basic idea in streamflow studies is to establish connections that generally exist, but attempts to identify such connections are largely dictated by the problem at hand and the system components in place. While numerous approaches have been proposed in the literature, our understanding of these connections remains far from adequate. The present study introduces the *theory of networks*, and in particular *complex networks*, to examine the connections in streamflow dynamics, with a particular focus on spatial connections. Monthly streamflow data observed over a period of 52 years from a large network of 639 monitoring stations in the contiguous United States are studied. The connections in this streamflow network are examined using the concept of *clustering coefficient*, which is a measure of local density and quantifies the network's tendency to cluster. The clustering coefficient analysis is performed with several different threshold levels, which are based on correlations in streamflow data between the stations. The clustering coefficient values of the 639 stations are used to obtain important information about the connections in the network and their extent, similarity and differences between stations/regions, and the influence of thresholds. The relationship of the clustering coefficient with the number of links/actual links in the network and the number of neighbors is also addressed. The results clearly indicate the usefulness of the network-based approach for examining connections in streamflow, with important implications for interpolation and extrapolation, classification of catchments, and predictions in ungaged basins.

1 Introduction

Streamflow forms an important input for a wide range of applications in hydrology, water resources, environment, and ecosystem. However, its estimation or prediction is an enormously challenging problem, since streamflow arises as a result of complex and nonlinear interactions between climate inputs (external factors) and landscape characteristics (internal factors) that occur over a wide range of spatial and temporal scales. For instance, streamflow is governed not only by the distribution of rainfall (in both space and time) but also by the nature and state of the catchment (e.g. topography, vegetation, soil, geology); see Beven (2006) for a compilation of, and stimulating insight into, some early "benchmark" studies (1933–1984) on streamflow generation processes. Attempts to monitor, model, and predict streamflow have been a central topic in hydrology during the last century or so; see, for example, Salas et al. (1995), Grayson and Blöschl (2000), Duan et al. (2003), Mishra and Coulibaly (2009), and Hrachowitz et al. (2013) for comprehensive accounts on streamflow monitoring, modeling, and prediction.

Despite their efforts and contributions, studies on streamflow have and continue to encounter at least two major challenges: (1) determination of the locations, number, and density of streamflow gaging stations for monitoring data and representation of process variability; and (2) identification of the appropriate scientific concepts and mathematical techniques/models for a more solid conceptual understanding of the catchment systems, proper analysis of the data, and reliable interpretation of the outcomes. It is true that recent developments in measurement technology, computational power, and mathematical sophistication have generally played an important role in overcoming these challenges to a certain extent. It can also not be denied, however, that the same developments have, at times, played an indirect role in creating imbalance and hindering true progress, as they have contributed to the perhaps unnecessary complexification in models (rather than simplification), highly specialized conceptual notions that are often suitable only for specific situations (rather than generalization frameworks that suit all conditions), difficult-to-bridge gaps between theory and practice, and lack of communication among researchers as well as between researchers and practitioners; see, for example, Perrin et al. (2001), Beven (2002), Kirchner (2006), Sivakumar (2008), and Young and Ratto (2009) for some details.

HESSD
11, 7255–7289, 2014

It is important to recognize that a fundamental idea in streamflow (and other hydrologic) studies is to establish connections that generally exist between the different elements or items (known or assumed) of the underlying system. Depending upon the situation (e.g. catchment, purpose, problem), these elements include hydroclimatic variables, catchment characteristics, model parameters, and others (and their combinations), and their connections are often different with respect to space, time, and space-time. Unraveling the nature and extent of these connections has always been a great challenge, not to mention the challenge in the identification of all the relevant elements in the first place. Thus far, a plethora of concepts and methods has been proposed and applied for studying the connections associated with streamflow, including those based on time, distance, correlation, variability, scale, patterns, and many other properties/measures as well as their combinations and variants, in both single-variable and multi-variable perspectives; see, for example, Gupta et al. (1986), Salas et al. (1995), Grayson and Blöschl (2000), Yang et al. (2004), Archfield and Vogel (2010), and Li et al. (2012) for some details. Despite the progress made through these concepts and methods, our understanding of the connections in streamflow is still far from adequate.

In view of this, there is indeed a need to greatly advance our studies on streamflow connections. Some important current and foreseeable future problems, including our ever-increasing demands for water, the potential impacts of climate change on water security and hydroclimatic disasters, and the numerous issues associated with the management of our environment and ecosystems, further reflect the urgency to this need. A greater understanding of streamflow connections will also enhance our recent and current efforts in the estimation of data at ungaged locations (e.g. predictions in ungaged basins – PUB) (see Hrachowitz et al., 2013) and development of a generalization framework for hydrologic modeling (e.g. catchment classification) (see Sivakumar et al., 2014), among others. The question, however, remains on the identification of a suitable theory that can help bring advancement to studies on streamflow connections. In this regard, recent developments in the field of complex systems science can

HESSD
11, 7255–7289, 2014

offer some crucial clues. The present study introduces the theory of *complex networks*, or simply *networks*, for studying connections in streamflow. In particular, the study focuses on spatial connections in streamflow.

The origin of the concept of networks can be traced back to the works of Leonhard Euler, during the first half of the eighteenth century, on the Seven Bridges of Königsberg (Euler, 1741), which laid the foundations of what would become popularly known as *graph theory*. Graph theory witnessed several important theoretical developments in the nineteeth century, including *topology* (originally introduced as *topologie* in German) (Listing, 1848) and *trees* (Cayley, 1857). Further significant advances were made during the twentieth century, especially with the development of *random graph theory* by Erdös and Rényi (1960). The concepts of graph theory, and random graph theory in particular, have found a wide variety of applications in numerous fields, including linguistics, physics, chemistry, biology, sociology, engineering, economics, and ecology; see, for example, Berge (1962), Bondy and Murty (1976), and Bollobás (1998) for extensive reviews.

Despite the above-mentioned developments and applications, studies on graph theory, including random graph theory, had some major deficiencies. First, the studies largely focused on networks that are regular, simple, small, and static. As a result, they are generally unsuitable for examining real networks, as such networks are often highly irregular, complex, large, and dynamically evolving in time. Second, even while examining complex and large-scale networks, they assumed that such networks are wired randomly together (Erdös and Rényi, 1960). Such an assumption, however, is not necessarily valid for real networks, since order and determinism are inherent in real systems and networks. Indeed, real networks are neither completely ordered nor completely random, but generally exhibit important properties of both. These observations motivated a renewed and fresh look of random graph theory towards the end of the last century (e.g. Watts and Strogatz, 1998; Barabási and Albert, 1999), and gave birth to a new movement of interest and research in studying real and complex networks, under the umbrella of the *new science of networks*. They also led to

HESSD
11, 7255–7289, 2014

new discoveries about complex networks, including *small-world networks* (Watts and Strogatz, 1998), *scale-free networks* (Barabási and Albert, 1999), *network motifs* (Milo et al., 2002), as well as other notable advances, such as a new method for identifying *community structure* (Girvan and Newman, 2002). Since then, the science of networks has found applications in many different fields, including natural and physical sciences, social sciences, medical sciences, economics, and engineering and technology (e.g. Albert et al., 1999; Bouchaud and Mézard, 2000; Newman, 2001; Liljeros et al., 2001; Tsonis and Roebber, 2004; Davis et al., 2013). In hydrology, applications of networks are just starting to emerge, and so far include river networks, virtual water trade, precipitation, and agricultural pollution due to international trade, among others (Rinaldo et al., 2006; Suweis et al., 2011; Dalin et al., 2012; Boers et al., 2013; Scarsoglio et al., 2013). In a very recent study, Sivakumar (2014) has argued that networks can be useful for studying all types of connections in hydrology and, hence, can provide a generic theory for hydrology.

With the encouraging results reported by the above studies, the present study explores the usefulness of the theory of networks for studying connections in streamflow, especially the spatial connections. To this end, monthly streamflow data observed over a period of 52 years (1951–2002) from each of 639 gaging stations in the contiguous United States are studied. The connections are examined using the concept of *clustering coefficient*. The clustering coefficient is a measure of local density and, hence, quantifies the tendency of a network to cluster. The implications of the clustering coefficient results for interpolation/extrapolation of streamflow data as well as for classification of catchments are also discussed.

The rest of this paper is organized as follows. Section 2 introduces the concept of networks and describes the procedure for calculation of the clustering coefficient in a network. Section 3 presents details of the study area and streamflow data considered. Section 4 reports the results, first from the traditional linear correlation analysis and then from the network-based clustering coefficient analysis. Section 5 highlights the implications of the results.

HESSD
11, 7255–7289, 2014

2 Network and clustering coefficient

2.1 Network

A *network* or a *graph* is a set of points connected together by a set of lines, as shown in Fig. 1. The points are referred to as *vertices* or *nodes* and the lines are referred to as *edges* or *links*; here, the term *nodes* are used for points and the term *links* are used for lines. Mathematically, a network can be represented as $G = \{P, E\}$, where P is a set of N nodes ($P_1, P_2, \ldots, P_N$) and E is a set of n links. The network shown in Fig. 1 has $N = 7$ (nodes) and $n = 8$ (links), with $P = \{1,2,3,4,5,6,7\}$ and $E = \{\{1,7\},\{2,3\},\{2,5\},\{2,7\},\{3,7\},\{4,7\},\{5,6\},\{6,7\}\}$.

Figure 1 is perhaps the simplest form of network, i.e. one with a set of identical nodes connected by identical links. There are, however, many ways in which networks may be more complex. For instance, a network: (1) may have more than one different type of node and/or link, (2) may contain nodes and links with a variety of properties, such as different weights for different nodes and links depending on the strength of nodes and connections, (3) may have links that can be directed (pointing in only one direction), with either cyclic (i.e. containing closed loops of links) or acyclic form, (4) may have multilinks (i.e. repeated links between the same pair of nodes), self-links (i.e. links connecting a node to itself), and hyperlinks (i.e. links connecting more than two nodes together); and (5) may be bipartite, i.e. containing nodes of two distinct types, with links running only between unlike types.

There are many different ways and measures to study the characteristics of networks. In the context of the modern theory of *complex networks* (which also include random graphs), three concepts are prominent: (1) clustering coefficient, (2) small-world networks; and (3) degree distribution. As the present study uses the concept of clustering coefficient for studying streamflow connections, it is described next.

HESSD
11, 7255–7289, 2014

B. Sivakumar and F. M. Woldemeskel

2.2 Clustering coefficient

The clustering coefficient quantifies the tendency of a network to cluster, which is one of the most fundamental properties of networks (Watts and Strogatz, 1998). The clustering coefficient of a network is basically a measure of local density. The concept of clustering has its origin in sociology, under the name "fraction of transitive triples" (Wasserman and Faust, 1994). The procedure for calculating the clustering coefficient is as follows.

Let us consider first a selected node i in the network, having k_i links which connect it to k_i other nodes. For illustration, Fig. 2 presents a network consisting of eight nodes, with the node i having four links (see Fig. 2, left). The four nodes corresponding to these four links are the *neighbors* of node i; the neighbors are identified based on some conditions (e.g. correlation between node i and other nodes in the network). If the neighbors of the original node (i) were part of a cluster, there would be $k_i(k_i-1)/2$ links between them. As shown in Fig. 2 (right), there are $4(4-1)/2=6$ links in the *cluster* of node i. The clustering coefficient of node i is then given by the ratio between the number E_i of links that actually exist between these k_i nodes (shown as solid lines on Fig. 2, right) and the total number $k_i(k_i-1)/2$ (i.e. all lines on Fig. 2, right),

$$C_i = \frac{2E_i}{k_i(k_i-1)} \tag{1}$$

The clustering coefficient of the whole network C is the average of the clustering coefficients C_i's of all the individual nodes.

The clustering coefficient of a random graph is $C=p$ (where p is the probability of two nodes being connected), since the links in a random graph are distributed randomly. However, the clustering coefficient of real networks is generally much larger than that of a comparable random network (i.e. having the same number of nodes and links as the real network). Therefore, the clustering coefficient analysis offers useful information about the nature of the network and, hence, the appropriate model (e.g. level of complexity), among others.

HESSD
11, 7255–7289, 2014

B. Sivakumar and F. M. Woldemeskel

3 Study area and data

In the present study, streamflow data from the United States are studied to explore the usefulness of the theory of networks for identifying connections in streamflow, with a focus on spatial connections. Monthly data from an extensive network of 639 streamflow gaging stations in the contiguous US are studied. The locations of these 639 stations are shown in Fig. 3. The streamflow data are obtained from the US Geological Survey database (http://nwis.waterdata.usgs.gov/nwis). Streamflow data in the US are commonly expressed in “water years,” which commence in October. The data used in this study are those observed over a period of 52 years (October 1951–September 2003), and are average monthly values.

During the past few decades, a large number of studies have investigated the above streamflow dataset (or a part or variant of it) in many different contexts (e.g. Slack and Landwehr, 1992; Kahya and Dracup, 1993; Tootle and Piechota, 2006; Sivakumar and Singh, 2012). Some of these studies have explicitly addressed the connections of streamflow, although with large-scale climatic patterns and relevant indices, including El-Niño, La-Niña, Southern Oscillation Index (SOI), Pacific North America (PNA) Index, and Pacific Decadal Oscillation (PDO). However, within the specific context of the network analysis for connections among streamflow stations presented here, as well as in the broader context of complex systems science for streamflow analysis, the study by Sivakumar and Singh (2012) is worth mentioning, as it has addressed the aspects of streamflow variability, nonlinearity, and dominant governing mechanisms, especially for studies on model simplification, data interpolation/extrapolation, and catchment classification framework.

The above 639 streamflow stations and the observed streamflow data exhibit tremendous variations in their characteristics, often by about four orders of magnitude. For instance: (1) basin drainage area ranges from 10.62 km^2 (4.1 mi^2) to 35 224 km^2 (13 600 mi^2), (2) station elevation ranges from 0 m to 2996 m (9830 ft), (3) mean flow ranges from 0.0549 m^3 s^{-1} (1.94 ft^3 s^{-1}) to 381.59 m^3 s^{-1} (13 476 ft^3 s^{-1}), (4) maximum

HESSD
11, 7255–7289, 2014

Complex networks for streamflow dynamics

B. Sivakumar and F. M. Woldemeskel

flow ranges from 0.878 $m^3 s^{-1}$ (31 $ft^3 s^{-1}$) to 2489 $m^3 s^{-1}$ (87 900 $ft^3 s^{-1}$); and (5) number of zero-flow months ranges from none to 424. Table 1 presents a summary of the minimum and maximum values of some important characteristics of the stations and flows, including the corresponding station numbers. Figure 3 presents the variations in the mean (Fig. 3a), standard deviation (Fig. 3b), and coefficient of variation (Fig. 3c) of flow values in all the 639 stations. Some important observations are:

- more than half of the 639 stations (340 stations) are small- to medium-sized basins, i.e. having a drainage area of less than 1000 km^2 (or approximately 400 mi^2);
- 137 stations have zero flows at least for one month, and the remaining 502 stations always have some flows every month;
- about half of the stations (49 %) have monthly mean flows of less than 10 $m^3 s^{-1}$ (approximately 350 $ft^3 s^{-1}$), while about 4 % of the stations have mean flows more than 100 $m^3 s^{-1}$ (approximately 3530 $ft^3 s^{-1}$). Similar observations are also made for standard deviation, with about 47 % and 4 %, respectively; and
- half of the stations (50 %) have CV values of flow less than 1.0, with most of the stations having values 0.5–1.0 are in the east, northeast, and northwest (see Fig. 3c). Of the remaining half, about one-fifth of the stations (9 % of the total) have CV values above as high as 2.0.

4 Analysis and results

The usefulness of the theory of networks for studying connections in streamflow is examined through the clustering coefficient analysis on the monthly streamflow data from the above 639 stations in the United States. To put the clustering coefficient analysis in a proper perspective, a preliminary linear correlation-based analysis is also performed.

HESSD
11, 7255–7289, 2014

4.1 Correlation analysis

A common approach to examine connections between streamflow observed at different stations is through a simple linear cross correlation analysis, where the correlation for any given station is given by the average of its correlation with all the other stations. Several variants of this procedure are also usually considered. These include: *nearest* neighbors – for example, *number of nearby stations based on distance* or stations within a pre-defined *region of geographic promixity* or *neighborhood*, with equal or unequal weightage (e.g. inverse distance); and *similar* stations – stations with *similar* properties (e.g. in terms of climate, rainfall, basin characteristics, land use), which may or may not include nearest stations. These and many other *correlation-based* procedures (e.g. spline fitting) are routinely employed for interpolation and extrapolation of streamflow and other hydrologic data.

In this study, two of the above-mentioned procedures are employed for examining the monthly streamflow from the 639 stations: (1) for each station, the correlation is the average of its correlation with all the other 638 stations; and (2) for each station, the correlation is the average of correlations for a *certain number of nearest neighbors* – 30, 15, and 5 neighbors. When all the 638 stations are considered, the correlation values are generally very low, as expected, with only 0.5 % of the stations exceeding a value of 0.4 (see Fig. 4a). This is mainly due to the consideration of a very large region, with the stations coming from different climatic, catchment, land use, and other characteristics. When the number of stations is reduced, the results get generally better – see Fig. 4b (30 neighbors), Fig. 4c (15 neighbors), and Fig. 4d (5 neighbors). Among the three neighborhood cases, the best correlation results are obtained when the neighborhood is the smallest, i.e. 5 neighbors (Fig. 4d), with a large number of stations having correlations above 0.7.

While one can study a large number of combinations in terms of the *neighborhood*, what is evident from even the very few cases presented here is that there are obvious *regional* patterns in terms of correlations, regardless of the number of neighbors. These

HESSD
11, 7255–7289, 2014

regional patterns are considered to have important implications for a wide range of studies in hydrology and water resources, as they are commonly used as a basis for interpolation and extrapolation of streamflow and, subsequently, for water resources assessment, planning, and management. However, as Sivakumar and Singh (2012) point out, through their nonlinear dynamic study on streamflow data from the western United States, the use of *regional* patterns as basis for streamflow studies may be misleading, as such patterns are not necessarily a true representation of the actual connections between the stations but may just be spurious. The obvious question, therefore, is: how to identify if the connections are actual or spurious? This is where the ideas from the theory of networks can be particularly useful, as presented next using the clustering coefficient analysis of the streamflow data from the 639 stations.

4.2 Network analysis – clustering coefficient

The clustering coefficient is calculated for the monthly streamflow data from the network of 639 stations in the United States, according to the procedure described in Sect. 2. The essence of the procedure for the streamflow data is as follows. For a given streamflow station or node i, the nearest neighbors k_i in the network of 639 stations (more specifically, the remaining 638 stations) are identified based on a (pre-specified) threshold value (T). To define the threshold value, the correlations in streamflow data between different stations are considered as a reasonable measure. With this, if, for example, the correlation between station i and any other station(s) in the entire network of 639 stations exceeds the threshold value, then that station(s) is considered as a *neighbor(s)*, k_i, for station i. The *cluster* of these k_i neighbors then forms the basis for identifying the *actual connections*. Therefore, the *actual connections* are those links in the *cluster* of stations (not just *nearest* stations) having correlations among themselves exceeding the threshold value.

In this study, several different threshold values are considered for calculation of the clustering coefficient. Although there are no definitive guidelines for selection of the threshold values for streamflow (and other hydrologic) data, our experience in streamflow studies, especially spatial and temporal correlations, offers some useful clues. For instance, streamflow data generally exhibit high spatial correlations (when compared to rainfall values, for example), especially at the monthly scale. With this knowledge, and also with the condition that $-1 < T < 1.0$, closer intervals of values are considered at the higher end of correlations and vice-versa. In addition, very low values (say, $T < 0.30$) and very high values (say, $T > 0.85$) do not offer much help in the analysis; for instance, $T < 0.30$ normally results in a very large number of neighbors, while $T > 0.85$ results in a very small number. Considering all these, eight threshold values are used for analysis: 0.30, 0.40, 0.50, 0.60, 0.70, 0.75, 0.80, and 0.85.

Figure 5a–d shows the clustering coefficient values obtained for the 639 stations for four different threshold values: 0.70, 0.75, 0.80, and 0.85; in these plots, for better illustration, the clustering coefficient values are grouped into five different ranges. Table 2 presents the number of stations falling under different ranges of clustering coefficient values. From an overall perspective, the clustering coefficient results indicate certain similarity at some stations/regions but significant differences at others. They also offer some specific observations:

- Even *nearest* stations have significantly different characteristics (e.g. connections), as part of a network. Some stations have very strong connections, while others have almost no or only very weak connections. For instance, the few geographically closer stations in Florida in the southeast region (see Fig. 5a–d) are an excellent example. Regardless of the threshold, these few stations have clustering coefficient values varying anywhere from 0 to 1.0.

- Even *distant* stations have significantly similar characteristics, i.e. they have very strong (or very weak) connections as part of a network. The similar (very high or very low) clustering coefficient values obtained for a number of stations all across the United States, regardless of their geographic promixity, offer evidence to this; for example, regardless of the threshold value, the green circles (see Fig. 5a–d), representing the clustering coefficient range 0.8–1.0, are present all over the

HESSD
11, 7255–7289, 2014

United States, northwest to southwest to midwest to northeast to southeast. Similar observations are made also for other clustering coefficient ranges, for one or more threshold values; see the deep pink circles (C_i = 0.60–0.79) and blue circles (C_i = 0);

– There are significant changes in characteristics with respect to the threshold values. For instance, as can be seen from Fig. 5 and Table 2, for threshold values of 0.7 and 0.85, the number of stations falling within the clustering coefficient range of 0.9–1.0 is 84 and 156, respectively, whereas that falling within the clustering coefficient range of 0.7–0.8 is 149 and 80, respectively; and

– Although there are changes in the number of stations having similar clustering coefficient values with respect to thresholds, there is no consistency in the trend of changes.

While the usefulness of the clustering coefficient values in assessing connections between streamflow stations and identifying regions having similarity/differences is abundantly clear, the *actual links* in the network would certainly offer more specific details as to where and how connections exist. To facilitate this, Fig. 6 shows the *actual links* for four selected streamflow stations (red circles) for threshold values of 0.75 (Fig. 6a), 0.80 (Fig. 6b), and 0.85 (Fig. 6c); the nodes and links for T = 0.70 are too many, and so do not offer a good visualization. In each of these plots, for the station of interest (red circle), a green circle indicates a station that has a correlation coefficient value exceeding the threshold, and a black circle indicates a station that has a correlation coefficient value smaller than the threshold. The lines are the *actual links* among all the links available for the *cluster of neighbors* (green circles only). The plots clearly indicate which stations are *actually* connected to which other. The plots make it abundantly clear that *geographic promixity* does not always result in greater correlation, and the *actual links* can go for large distances. Among the various observations that can be made, the ones for the two stations in the northwest are certainly interesting. Despite being in the same region, the two stations exhibit significantly different connectivity

HESSD
11, 7255–7289, 2014

characteristics, for example, for threshold level 0.85 (Fig. 6b), with one showing all the actual connections within a small neighborhood (see the enlarged plot on the top left) while the other showing no clear neighborhood for connectivity (see the enlarged plot on the bottom left). The latter station (see bottom left) is an even more curious case, as most of the neighbors of this station seem to be beyond its (perceived) *circle of geographic influence*. The actual links observed for the other threshold values also support the above observations.

These observations clearly suggest that our usual approach with consideration of geographic proximity, nearest neighbors, regional patterns, and linear correlation-based techniques for studying connections in streamflow may have serious limitations. Clustering coefficient, and other network-based techniques, offers a better means to examine streamflow connections. In what follows, we explore the clustering coefficient results even further.

As the clustering coefficient of a network is based on the *actual links* among *all links* in the *cluster* of neighbors of a node (rather than just the links between a node and its neighbors), it would be interesting to see how it changes with respect to *all links* and *actual links*. To this end, Fig. 7a–d shows the clustering coefficient values against the *number of all links* (red circles) and the *number of actual links* (blue circles) for threshold values of 0.70, 0.75, 0.80, and 0.85 for the monthly streamflow data from the United States. The results lead to the following major observations:

- in general, regardless of the threshold value, there is an inverse relationship between the clustering coefficient and number of links (both for *all links* and *actual links*), i.e. higher clustering coefficient for smaller number of links and vice-versa;

- the inverse relationship between the clustering coefficient and number of links is generally more evident for lower thresholds (see Fig. 7a and b) when compared to higher thresholds (see Fig. 7c and d). When the threshold is very high ($T = 0.85$), this relationship seems to cease to exist;

HESSD
11, 7255–7289, 2014

- the clustering coefficient is generally far more sensitive when the number of links is smaller (see the significant larger spread of circles on the Y-axis), but has only very little or almost no sensitivity for a larger number of links (see the very narrow spread followed by a tapering towards a fixed value – especially in Fig. 7a and b). Further, larger numbers of links almost always give lower clustering coefficients; and

- for a given number of links, the clustering coefficient for a lower threshold is generally higher than that for a higher threshold.

Another useful way to look at the clustering coefficient of a network is its relationship with the number of neighbors (k_i), which is defined by the threshold value and dictates the (number of) links and actual links. Figure 8a–d shows the relationship between the clustering coefficient values and the number of neighbors for threshold values of 0.70, 0.75, 0.80, and 0.85 for the monthly streamflow data. The results generally indicate an inverse relationship between the clustering coefficient and number of neighbors, but such a relationship is far more evident for lower threshold values (see Fig. 8a and b) than that for higher threshold values (see Fig. 8c and d). Again, the clustering coefficient is generally far more sensitive when the number of neighbors is smaller (see the larger spread towards the left), but becomes less sensitive for a larger number of neighbors (see the narrow spread towards the right). These observations are somewhat consistent with those made in regard to the number of links (Fig. 7). It is important to recall, however, that the neighbors are not necessarily geographic but defined by the threshold values (as shown in Fig. 6).

While these results and observations are still preliminary in nature, they seem to suggest that there is a particular threshold value or range beyond which the inverse relationship between the clustering coefficient and number of neighbors/links/actual links in the streamflow network may not hold well for monthly streamflow data from the United States, and streamflow data in general.

HESSD
11, 7255–7289, 2014

B. Sivakumar and F. M. Woldemeskel

Title Page
Abstract | Introduction
Conclusions | References
Tables | Figures
◀| |▶
◀ ▶
Back | Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

Finally, the question arises as to the type of network. As mentioned previously, the clustering coefficient of a whole network (C) is the average of the clustering coefficients C_i's of all the individual nodes. The clustering coefficient of the eight different networks of the above 639 streamflow stations corresponding to threshold values of 0.30, 0.40, 0.50, 0.60, 0.70, 0.75, 0.80, and 0.85 is 0.79, 0.77, 0.73, 0.71, 0.70, 0.70, 0.68, and 0.67 (see Table 2). These generally high clustering coefficient values seem to suggest that the streamflow monitoring network of 639 stations is not a random graph, since a (comparable) random graph, where the links are distributed randomly, will have a typically very low clustering coefficient, i.e. $C = p$, where p is the probability of two nodes being connected. As (natural) streamflow dynamics are neither completely random (there are inherent deterministic patterns) nor completely ordered (there are inherent stochastic components) (see Sivakumar, 2011; Sivakumar and Singh, 2012 for some details), it is also reasonable to assume that streamflow networks are not random graphs, but networks of some other nature. Whether they are *small-world* or *scale-free* or other types of networks remains to be seen. Studies in this direction are currently underway, details of which will be reported in the future.

5 Study implications

One of the basic requirements in studying streamflow dynamics is to identify connections in space or time or space-time, depending upon the purpose. Although a wide variety of approaches have been developed and applied to identify connections in streamflow dynamics, there is no question that significant improvements are still needed. In this regard, modern developments in the field of network theory, especially complex networks, offer new avenues, both for their generality about systems and for their holistic perspective about connections.

The present study has made an initial attempt to apply the ideas developed in the field of complex networks to examine connections in streamflow dynamics, with particular focus on spatial connections. Application of the concept of clustering coefficient,

HESSD
11, 7255–7289, 2014

which is a measure of local density and quantifies the tendency of a network to cluster, to monthly streamflow data from a large network of 639 monitoring stations in the contiguous United States has offered some very interesting results. The clustering coefficient values for the 639 stations suggest that: (1) even nearest stations can have significantly different connections and distant stations can have significantly similar connections, (2) connections can be significantly different for different threshold levels, (3) there is generally an inverse relationship between the clustering coefficient and number of neighbors, number of all links, and actual links (in the cluster of neighbors), (4) the clustering coefficient is far more sensitive when the number of neighbors/number of links is smaller, but has only little or no sensitivity when the latter is larger; and (5) the high clustering coefficient value obtained for the entire network is not consistent with the one expected for a random graph, suggesting that the streamflow network is likely to be small-world or scale-free or some other type.

Although the present results are preliminary, they offer important information about the connections that possibly exist in the streamflow network, and especially their extent. The clustering coefficient values, and the *actual links*, are particularly useful in the identification of the specific regions where interpolation and extrapolation of streamflow data may be more effective and also of the specific stations whose data can be more reliable for such purposes. For instance, regions consisting of stations with high clustering coefficient values would generally provide a more accurate estimation of streamflow when interpolation and extrapolation schemes are employed. It is also important to emphasize, however, that such a region is identified based on *cluster of actual connections*, rather than based on our traditional way of geographic proximity, nearest neighbors, regional patterns, and linear correlations. The clustering coefficient values can also offer important clues and guidelines as to the setting up/removal of streamflow monitoring stations in a region. For instance, if a region consists of stations with very high clustering coefficients, then installing additional monitoring stations will not offer any significant benefits. Indeed, one or more monitoring stations from such a region may be removed and the resources can be used in regions where additional

HESSD
11, 7255–7289, 2014

stations might offer greater benefits (e.g. in regions where the clustering coefficient values are low).

Finally, the present study and the results obtained have important implications for a wide range of issues and associated efforts in streamflow modeling, and hydrologic modeling in general. Among these are: (1) predictions in ungaged basins (PUB), where approaches based on nearest neighbors, regionalization, similarity, and other concepts are commonly adopted, (2) formulation of a catchment classification framework, for simplification and generalization in our modeling paradigm and better communication among/between researchers and practitioners; and (3) development of an integrated framework for water planning and management, including in studies on climate change impacts on water resources, that involves proper consideration and inclusion of stakeholders and concepts from a vast number of disciplines, including climate, hydrology, engineering, environment, ecology, social sciences, political sciences, economics, and psychology. In view of these, ideas gained from the modern theory of complex networks, and network theory at large, seem to have immense potential in hydrology and water resources.

Acknowledgements. Support for this work was provided by the Australian Research Council (ARC). Bellie Sivakumar acknowledges the financial support from ARC through the Future Fellowship grant (FT110100328).

References

Albert, R., Jeong, H., and Barabasi, A.-L.: Internet: diameter of the world wide web, Nature, 401, 130–131, 1999.

Archfield, S. A. and Vogel, R. M.: Map correlation method: selection of a reference streamgage to estimate daily streamflow at ungaged catchments, Water Resour. Res., 46, W10513, doi:10.1029/2009WR008481, 2010.

Barabási, A.-L. and Albert, R.: Emergence of scaling in random networks, Science, 286, 509–512, 1999.

Berge, C.: The Theory of Graphs and Its Applications, Matheun, Ann Arbor, MI, USA, 1962.

Beven, K. J.: Uncertainty and the detection of structural change in models of environmental systems, in: Environmental Foresight and Models: a Manifesto, edited by: Beck, M. B., Elsevier Science Ltd, Oxford, UK, 227–250, 2002.
Beven, K. J.: Benchmark Papers in Streamflow Generation Processes, IAHS Press, Wallingford, UK, 2006.
Boers, N., Bookhagen, B., Marwan, N., Kurths, J., and Marengo, J.: Complex networks identify spatial patterns of extreme rainfall events of the South American Monsoon System, Geophys. Res. Lett., 40, 4386–4392, doi:10.1002/grl.50681, 2013.
Bollobás, B.: Modern Graph Theory, Springer, New York, USA, 1998.
Bondy, J. A. and Murty, U. S. R.: Graph Theory with Applications, Elsevier Science Ltd, New York, USA, 1976.
Bouchaud, J.-P. and Mézard, M.: Wealth condensation in a simple model of economy, Physica A, 282, 536–540, 2000.
Cayley, A.: On the theory of the analytical forms called trees, Philos. Mag., 13, 172–176, 1857.
Dalin, C., Konar, M., Hanasaki, N., Rinaldo, A., and Rodriguez-Iturbe, I.: Evolution of the global virtual water trade network, P. Natl. Acad. Sci. USA, 109, 5989–5994, 2012.
Davis, K. F., D'Odorico, P., Laio, F., and Ridolfi, L.: Global spatio-temporal patterns in human migration: a complex network perspective, PLoS ONE, 8, e53723, doi:10.1371/journal.pone.0053723, 2013.
Duan, Q., Gupta, H. V., Sorooshian, S., Rousseau, A. N., Turcotte, R.: Calibration of Watershed Models, Water Science and Application Series, vol. 6, American Geophysical Union, Washington, DC, USA, 2003.
Erdös, P. and Rényi, A.: On the evolution of random graphs, Publ. Math. Inst. Hung. Acad. Sci., 5, 17–61, 1960.
Euler, L.: Solutio problematis ad geometriam situs pertinentis, Comment. Acad. Sci. Petropolitanae, 8, 128–140, 1741.
Girvan, M. and Newman, M. E. J.: Community structure in social and biological networks, P. Natl. Acad. Sci. USA, 99, 7821–7826, 2002.
Grayson, R. B. and Blöschl, G.: Spatial Patterns in Catchment Hydrology: Observations and Modeling, Cambridge University Press, Cambridge, UK, 2000.
Gupta, V. K., Rodriguez-Iturbe, I., and Wood, E. F.: Scale Problems in Hydrology: Runoff Generation and Basin Response, Water Science and Technology Library Series, Springer, Dordrecht, Holland, 1986.

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., Arheimer, B., Blume, T., Clark, M. P., Ehret, U., Fenicia, F., Freer, J. E., Gelfan, A., Gupta, H. V., Hughes, D. A., Hut, R. W., Montanari, A., Pande, S., Tetzlaff, D., Troch, P. A., Uhlenbrook, S., Wagener, T., Winsemius, H. C., Woods, R. A., Zehe, E., and Cudennec, C.: A decade of predictions in ungaged basins (PUB) – a review, Hydrolog. Sci. J., 58, 1198–1255, 2013.

Kahya, E. and Dracup, J. A.: US streamflow patterns in relation to the El Niño/Southern Oscillation, Water Resour. Res., 29, 2491–2503, 1993.

Kirchner, J. W.: Getting the right answers for the right reasons: linking measurements, analyses, and models to advance the science of hydrology, Water Resour. Res., 42, W03S04, doi:10.1029/2005WR004362, 2006.

Konar, M. and Caylor, K. K.: Virtual water trade and development in Africa, Hydrol. Earth Syst. Sci., 17, 3969–3982, doi:10.5194/hess-17-3969-2013, 2013.

Li, C., Singh, V. P., and Mishra, A. K.: Entropy theory-based criterion for hydrometric network evaluation and design: maximum information minimum redundancy, Water Resour. Res., 48, W05521, doi:10.1029/2011WR011251, 2012.

Liljeros, F., Edling, C., Amaral, L. N., Stanley, H. E., and Åberg, Y.: The web of human sexual contacts, Nature, 411, 907–908, 2001.

Listing, J. B.: Vorstudien zur Topologie, Vandenhoeck und Ruprecht, Göttingen, Germany, 811–875, 1848.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U.: Network motifs: simple building blocks of complex networks, Science, 298, 824–827, 2002.

Mishra, A. K. and Coulibaly, P.: Developments in hydrometric network design: a review, Rev. Geophys., 47, RG2001, doi:10.1029/2007RG000243, 2009.

Newman, M. E. J.: The structure of scientific collaboration networks, P. Nat. Acad. Sci. USA, 98, 404–409, 2001.

Perrin, C., Michel, C., and Andréassian, V.: Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments, J. Hydrol., 242, 275–301, 2001.

Rinaldo, A., Banavar, J. R., and Maritan, A.: Trees, networks, and hydrology, Water Resour. Res., 42, W06D07, doi:10.1029/2005WR004108, 2006.

Salas, J. D., Delleur, J. W., Yevjevich, V., and Lane, W. L.: Applied Modeling of Hydrologic Time Series, Water Resources Publications, Littleton, Colorado, USA, 1995.

HESSD
11, 7255–7289, 2014

Complex networks for streamflow dynamics

B. Sivakumar and F. M. Woldemeskel

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

Sivakumar, B.: Dominant processes concept, model simplification and classification framework in catchment hydrology, Stoch. Env. Res. Risk A., 22, 737–748, 2008.

Sivakumar, B.: Chaos theory for modeling environmental systems: philosophy and pragmatism, in: System Identification, Environmental Modelling, and Control System Design, edited by: Wang, L., and Garnier, H., Springer-Verlag, London, 533–555, 2011.

Sivakumar, B.: Networks: a generic theory for hydrology?, Stoch. Env. Res. Risk A., doi:10.1007/s00477-014-0902-7, in press, 2014.

Sivakumar, B. and Singh, V. P.: Hydrologic system complexity and nonlinear dynamic concepts for a catchment classification framework, Hydrol. Earth Syst. Sci., 16, 4119–4131, doi:10.5194/hess-16-4119-2012, 2012.

Sivakumar, B., Singh, V. P., Berndtsson, R., and Khan, S. K.: Catchment classification framework in hydrology: challenges and directions, J. Hydrol. Eng., A4014002, doi:10.1061/(ASCE)HE.1943-5584.0000837, 2014.

Slack, J. R. and Landwehr, V. M.: Hydro-Climatic Data Network (HCDN): a US Geological Survey Streamflow Data Set for the United States for the Study of Climate Variations, 1847–1988, US Geological Survey Open File Report 92–129, Reston, Virginia, USA, 1992.

Suweis, S., Konar, M., Dalin, C., Hanasaki, N., Rinaldo, A., and Rodriguez-Iturbe, I.: Structure and controls of the global virtual water trade network, Geophys. Res. Lett., 38, L10403, doi:10.1029/2011GL046837, 2011.

Tootle, G. A. and Piechota, T. C.: Relationships between Pacific and Atlantic ocean sea surface temperatures and US streamflow variability, Water Resour. Res., 42, W07411, doi:10.1029/2005WR004184, 2006.

Tsonis, A. A. and Roebber, P. J.: The architecture of the climate network, Physica A, 333, 497–504, 2004.

Wasserman, S. and Faust, K.: Social Network Analysis, Cambridge University Press, Cambridge, UK, 1994.

Watts, D. J. and Strogatz, S. H.: Collective dynamics of “small-world” networks, Nature, 393, 440–442, 1998.

HESSD

11, 7255–7289, 2014

Complex networks for streamflow dynamics

B. Sivakumar and F. M. Woldemeskel

Yang, D., Li, C., Hu, H., Lei, Z., Yang, S., Kusuda, T., Koike, T., and Musiake, K.: Analysis of water resources variability in the Yellow River of China during the last half century using historical data, Water Resour. Res., 40, W06502, doi:10.1029/2003WR002763, 2004.
Young, P. C. and Ratto, M.: A unified approach to environmental systems modeling, Stoch. Env. Res. Risk A., 23, 1037–1057, 2009.

HESSD
11, 7255–7289, 2014

Complex networks for streamflow dynamics

B. Sivakumar and F. M. Woldemeskel

Table 1. Characteristics of streamflow stations and data in the United States.

	Minimum	Maximum	Station
Drainage area (km^2)	10.62 ($4.1\,mi^2$)	35 224 ($13\,600\,mi^2$)	Minimum: #1188000 (CT) Maximum: #2226000 (GA)
Elevation (m)	0	2996 (9830 ft)	Minimum: #2310000 (FL) Maximum: #7083000 (CO)
Flow mean ($m^3\,s^{-1}$)	0.0549 ($1.938\,ft^3\,s^{-1}$)	381.59 ($13\,476\,ft^3\,s^{-1}$)	Minimum: #11063500 (CA) Maximum: #2226000 (GA)
Flow standard deviation ($m^3\,s^{-1}$)	0.110 ($3.888\,ft^3\,s^{-1}$)	373.75 ($13\,199\,ft^3\,s^{-1}$)	Minimum: #11063500 (CA) Maximum: #13317000 (ID)
Flow CV	0.11385	5.56342	Minimum: #6775500 (NE) Maximum: #6860000 (KS)
Flow skewness	0.45903	15.2588	Minimum: #6775500 (NE) Maximum: #6860000 (KS)
Flow Kurtosis	−0.33223	289.09	Minimum: #6775500 (NE) Maximum: #6860000 (KS)
Minimum flow ($m^3\,s^{-1}$)	0	63.91 ($2257\,ft^3\,s^{-1}$)	Minimum: 137 stations Maximum: #13317000 (ID)
Maximum flow ($m^3\,s^{-1}$)	0.878 ($31\,ft^3\,s^{-1}$)	2489 ($87\,900\,ft^3\,s^{-1}$)	Minimum: #1484500 (DE) Maximum: #6902000 (MO)
Number of zeros	0	424	Minimum: 502 stations Maximum: #10258500 (CA)

HESSD
11, 7255–7289, 2014

Complex networks for streamflow dynamics

B. Sivakumar and F. M. Woldemeskel

Table 2. Clustering coefficient values for monthly streamflow data from the United States.

Clustering coefficient range	Number of stations within each clustering coefficient range for threshold (T)							
	0.3	0.4	0.5	0.6	0.7	0.75	0.8	0.85
0.9–1.0	80	87	51	52	84	93	130	156
0.8–0.9	249	149	125	105	105	108	86	112
0.7–0.8	178	244	213	177	149	138	104	80
0.6–0.7	102	122	201	204	158	163	157	129
0.5–0.6	26	27	37	83	102	87	86	61
0.4–0.5	2	6	7	9	15	17	17	16
0.0–0.4	0	0	0	1	3	1	0	2
0.0	2	4	5	8	23	32	59	83
Entire Network	0.79	0.77	0.73	0.71	0.70	0.70	0.68	0.67

HESSD
11, 7255–7289, 2014

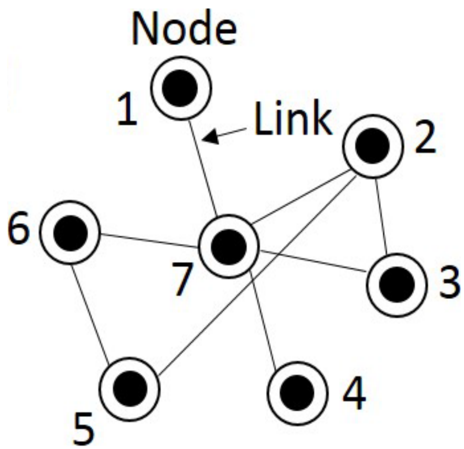

Figure 1. Network in its simplest form, i.e. an undirected network with only a single type of node and a single type of link.

HESSD
11, 7255–7289, 2014

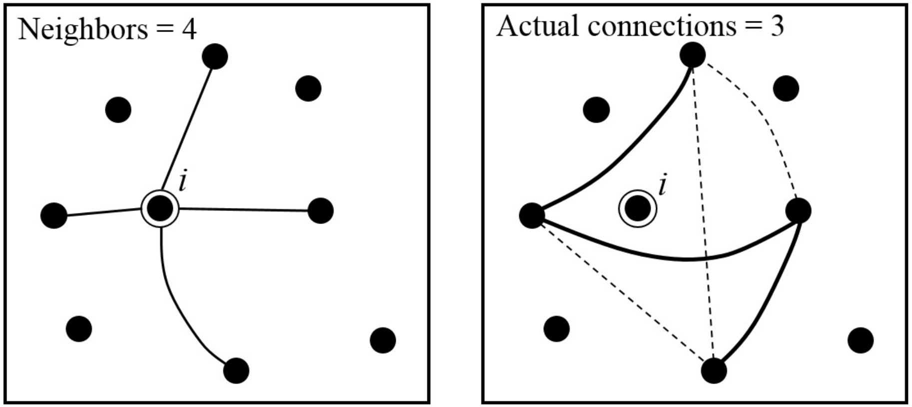

Figure 2. Connections in networks and calculation of clustering coefficient: nearest neighbors and actual connections.

HESSD
11, 7255–7289, 2014

Complex networks for streamflow dynamics

B. Sivakumar and F. M. Woldemeskel

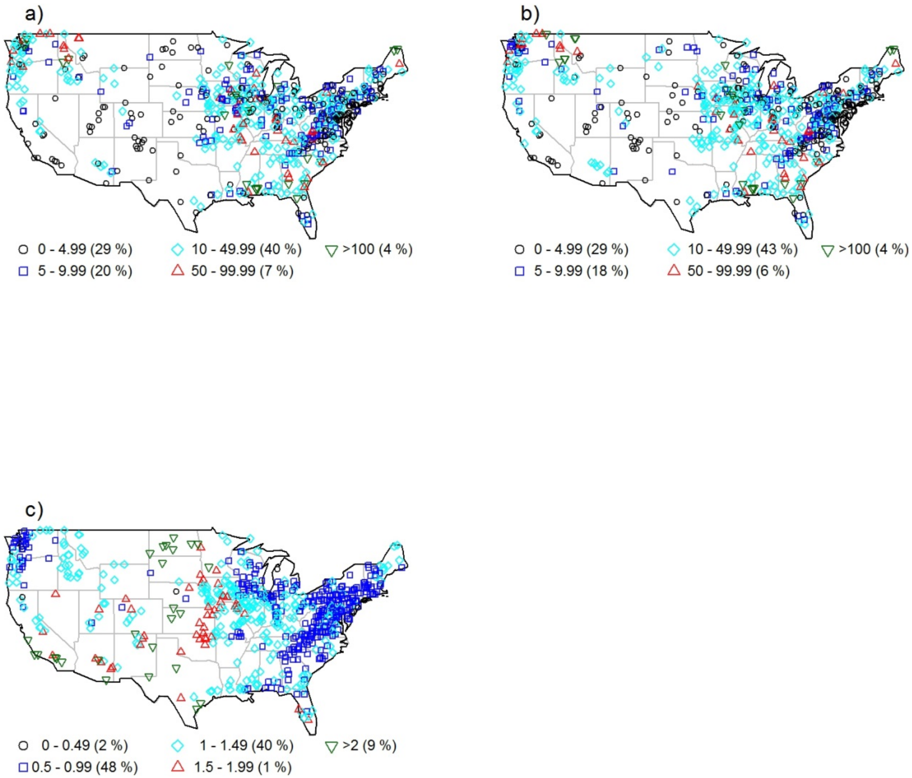

Figure 3. Characteristics of monthly streamflow observed at 639 stations in the United States: **(a)** mean; **(b)** standard deviation; and **(c)** coefficient of variation.

HESSD
11, 7255–7289, 2014
Complex networks for streamflow dynamics
B. Sivakumar and F. M. Woldemeskel

Title Page
Abstract | Introduction
Conclusions | References
Tables | Figures
Back | Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

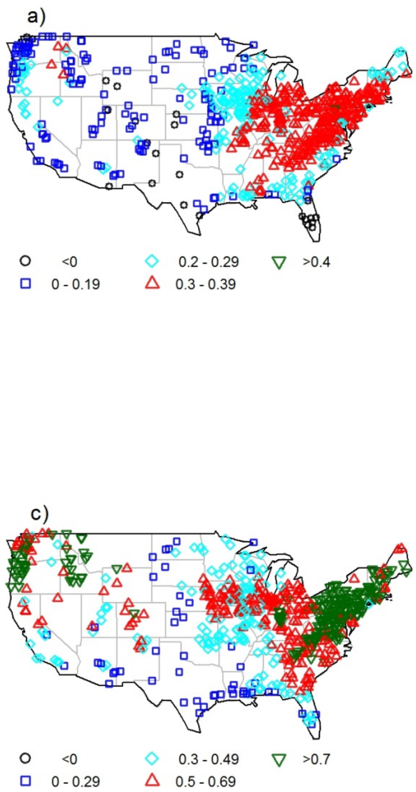

Figure 4. Linear correlation for streamflow: average of correlation with **(a)** all the 638 stations; **(b)** nearest 30 neighbors; **(c)** nearest 15 neighbors; and **(d)** nearest 5 neighbors.

HESSD
11, 7255–7289, 2014

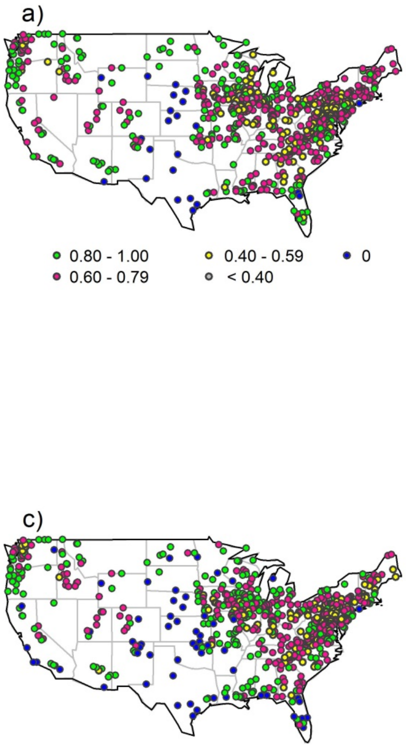

Figure 5. Clustering coefficients for four correlation thresholds: **(a)** 0.70; **(b)** 0.75; **(c)** 0.80; and **(d)** 0.85. The four ranges of 0.8–1.0, 0.6–0.8, 0.4–0.6, and < 0.4 are chosen for better visualization of results. See Table 2 for additional ranges.

HESSD
11, 7255–7289, 2014

Figure 6a. Links in streamflow network for threshold $T = 0.75$. Four nodes (stations) are chosen for better visualization.

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

HESSD

11, 7255–7289, 2014

Complex networks for streamflow dynamics

B. Sivakumar and F. M. Woldemeskel

Title Page
Abstract Introduction
Conclusions References
Tables Figures
◀I ▶I
◀ ▶
Back Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

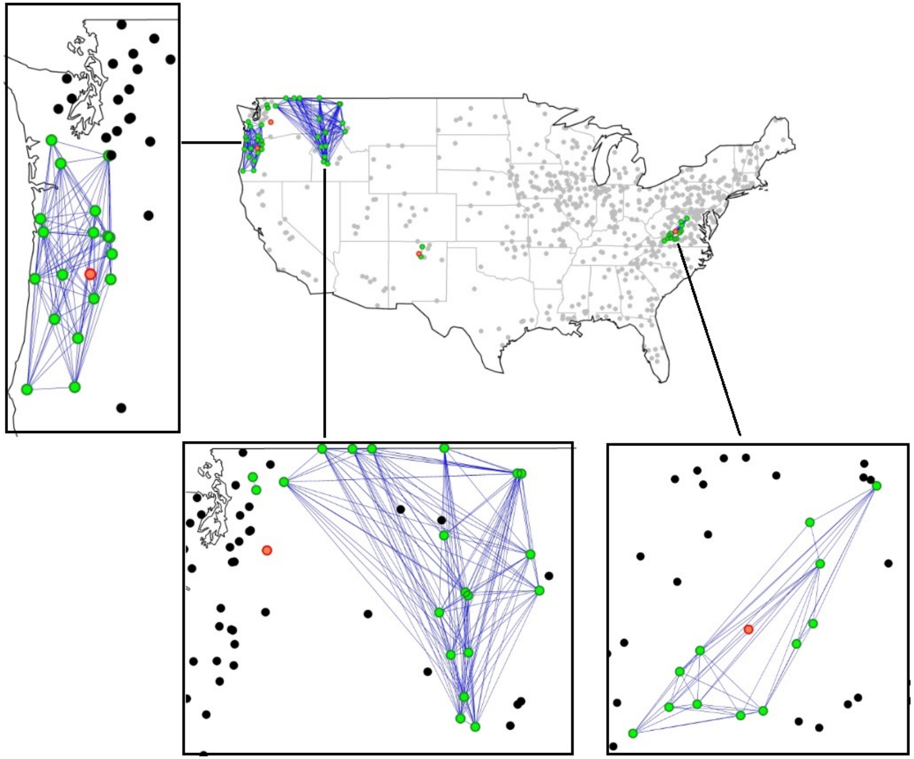

Figure 6b. Links in streamflow network for threshold $T = 0.80$. Four nodes (stations) are chosen for better visualization.

HESSD
11, 7255–7289, 2014

Complex networks for streamflow dynamics

B. Sivakumar and F. M. Woldemeskel

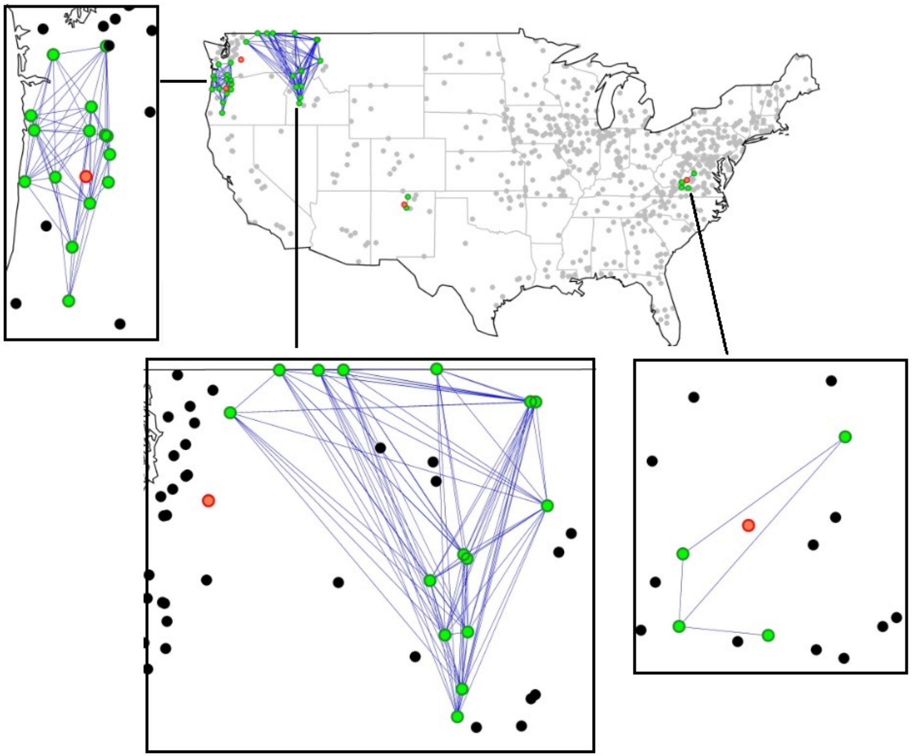

Figure 6c. Links in streamflow network for threshold $T = 0.85$. Four nodes (stations) are chosen for better visualization.

HESSD
11, 7255–7289, 2014

Complex networks for streamflow dynamics

B. Sivakumar and F. M. Woldemeskel

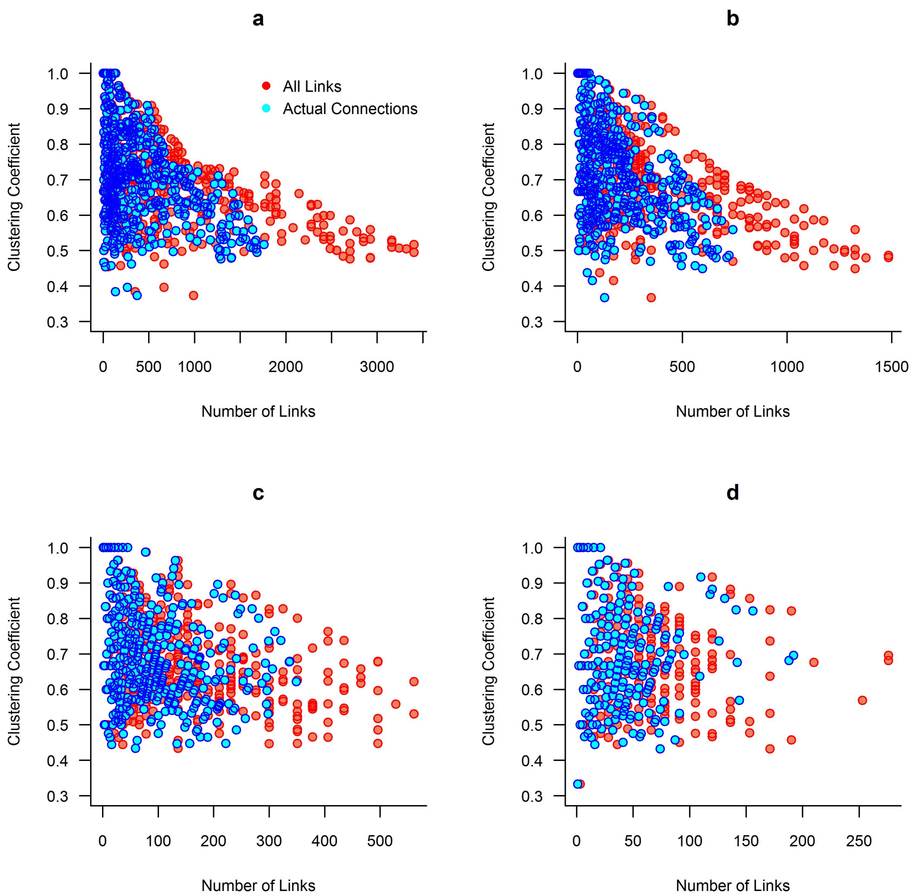

Figure 7. Relationship between clustering coefficient and number of links: **(a)** $T = 0.70$; **(b)** $T = 0.75$; **(c)** $T = 0.80$; and **(d)** $T = 0.85$. Both *all links* (red circles) and *actual links* (blue circles) are presented.

HESSD
11, 7255–7289, 2014

B. Sivakumar and F. M. Woldemeskel

HESSD
11, 7255–7289, 2014

Complex networks for streamflow dynamics

B. Sivakumar and F. M. Woldemeskel

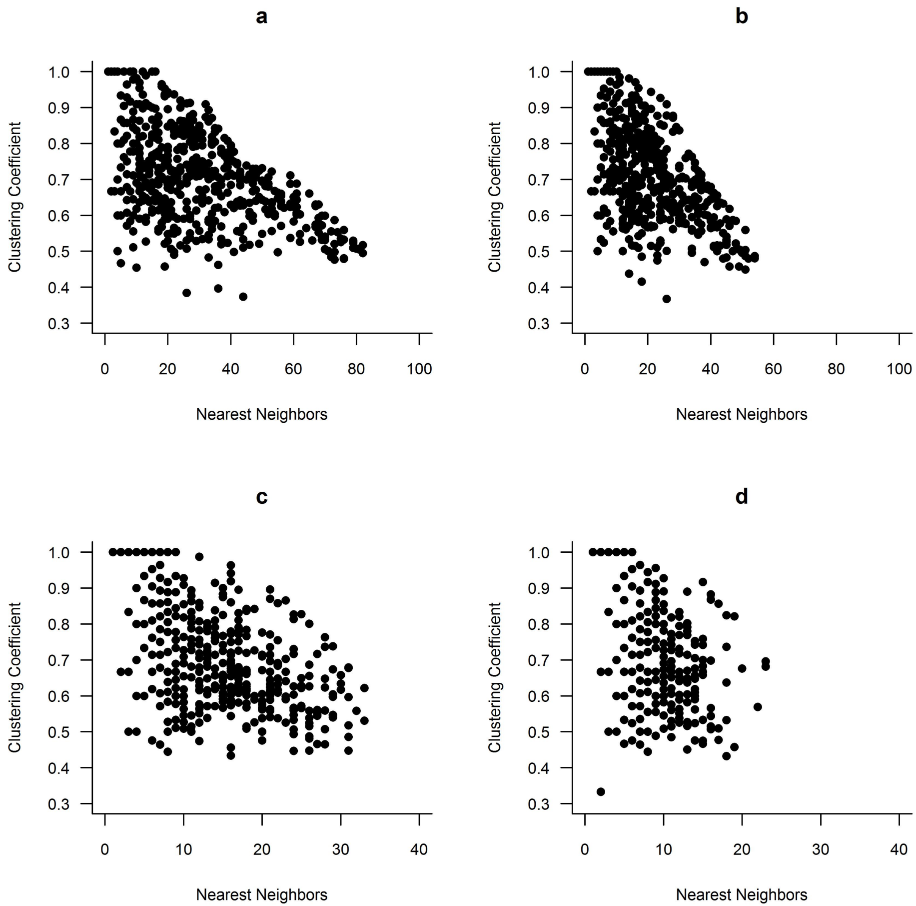

Figure 8. Relationship between clustering coefficient and number of nearest neighbors: **(a)** T = 0.70; **(b)** T = 0.75; **(c)** T = 0.80; and **(d)** T = 0.85.