Journal of Hydrology 579 (2019) 124185

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Reconstruction of daily rainfall data using the concepts of networks: Accounting for spatial connections in neighborhood selection

Shubham Tiwari[a], Sanjeev Kumar Jha[a,*], Bellie Sivakumar[b]

[a] Indian Institute of Science Education and Research Bhopal, Madhya Pradesh, India
[b] Department of Civil Engineering, Indian Institute of Technology Bombay, Powai, Mumbai 400076, India

ARTICLE INFO

This manuscript was handled by G. Syme

Keywords:
Reconstruction of rainfall
Complex network theory
Clustering coefficient
Inverse-distance-weighting interpolation
Nearest-neighborhood
Spatial connections

ABSTRACT

Accurate and reliable rainfall data is one of the fundamental prerequisites in hydrological modelling. The rainfall data at a desired location can be reconstructed using interpolation methods, such as Inverse Distance Weighting (IDW), which is frequently used in hydrology. In standard IDW neighbors are selected based on geographical proximity or nearest neighbor (IDW_NN). However, in a basin with variable topography, nearby rain gauges may be located at very different elevations and, thus, they may not accurately represent the spatial connection in rainfall. In this work, the theory of networks, with nodes and links as the basis, is applied to select neighbors while applying IDW. Two variants of neighbor selection models are proposed: IDW with linked neighbours (IDW_LN) and IDW with clustered neighbors (IDW_CN). For reconstruction, thirty years of daily rainfall data from 430 rain gauges in Murray-Darling Basin (MDB) are utilized. To evaluate the performance of the proposed models, one-station-leave-out cross validation approach is used and the associated Root-Mean-Squared-Error (RMSE) and Bias-percentage (BP) are calculated. Different values of number of neighbors (n), Correlation Threshold (CT) and Clustering Coefficient Range (CCR) are used to measure the errors associated with the proposed models. On comparing with IDW_NN, results show that reconstruction using IDW_LN has lower RMSE at about 30 percent of stations and lower BP for about 50 percent of stations; while IDW_CN shows lower RMSE at about 25 percent of stations and lower BP for about 45 percent of stations. The IDW_NN performed better than IDW_LN and IDW_CN at more than 50 percent of stations though the average error associated with all the three models are comparable for all CT values. In a natural system, a concept like traditional IDW (IDW_NN) may be more accurate than the network-based approach (IDW_LN and IDW_CN) but may not be completely efficient in accounting the spatial rainfall variability. The encouraging results for the reconstruction of rainfall in this study seem to indicate that the approach can be further helpful in the reconstruction of a wide range of meteorological parameters with spatial correlation.

1. Introduction

In hydrological modelling, reliable rainfall record is one of the basic prerequisites. The data collected from available rain gauges in a catchment may not provide the necessary or accurate information because of a number of reasons firstly, the density of the rain gauges may be high in certain areas but very low in other areas even within a catchment; secondly, some rain gauges may become dysfunctional and; thirdly, the topography may be complex having many inaccessible regions. The lack of rainfall data at a desired location can be overcome by using spatial interpolation. Previous studies have examined, among others, deterministic and stochastic approaches to reconstruct daily records using spatial interpolation algorithms ranging from simple techniques such as Thiessen polygons (Thiessen, 1911) or inverse distance weighting (Di Piazza et al., 2011) schemes to more complex and computationally intensive approaches such as geostatistical kriging (Buytaert et al., 2006) or regression based PRISM interpolation (Daly et al., 2008). Among all the different spatial interpolation methods, the inverse distance weighting (IDW) (Robertson, 1967) is perhaps the most commonly used method in hydrology.

The success of the IDW method depends fundamentally on the presence of positive spatial correlations between the data observed at neighboring rain gauges (Griffith, 1992). The underlying assumption is that the data from nearby points are more related than the data from locations far from each other, in accordance to Tobler's first law of Geography (Tobler, 1970). This presumption may not be applicable in certain situations, particularly in regions with complex topography (Shi et al., 2017). In such regions, even when rain gauges are located

* Corresponding author.
E-mail address: sanjeevj@iiserb.ac.in (S. Kumar Jha).

https://doi.org/10.1016/j.jhydrol.2019.124185
Received 8 May 2019; Received in revised form 25 September 2019; Accepted 26 September 2019
Available online 26 September 2019

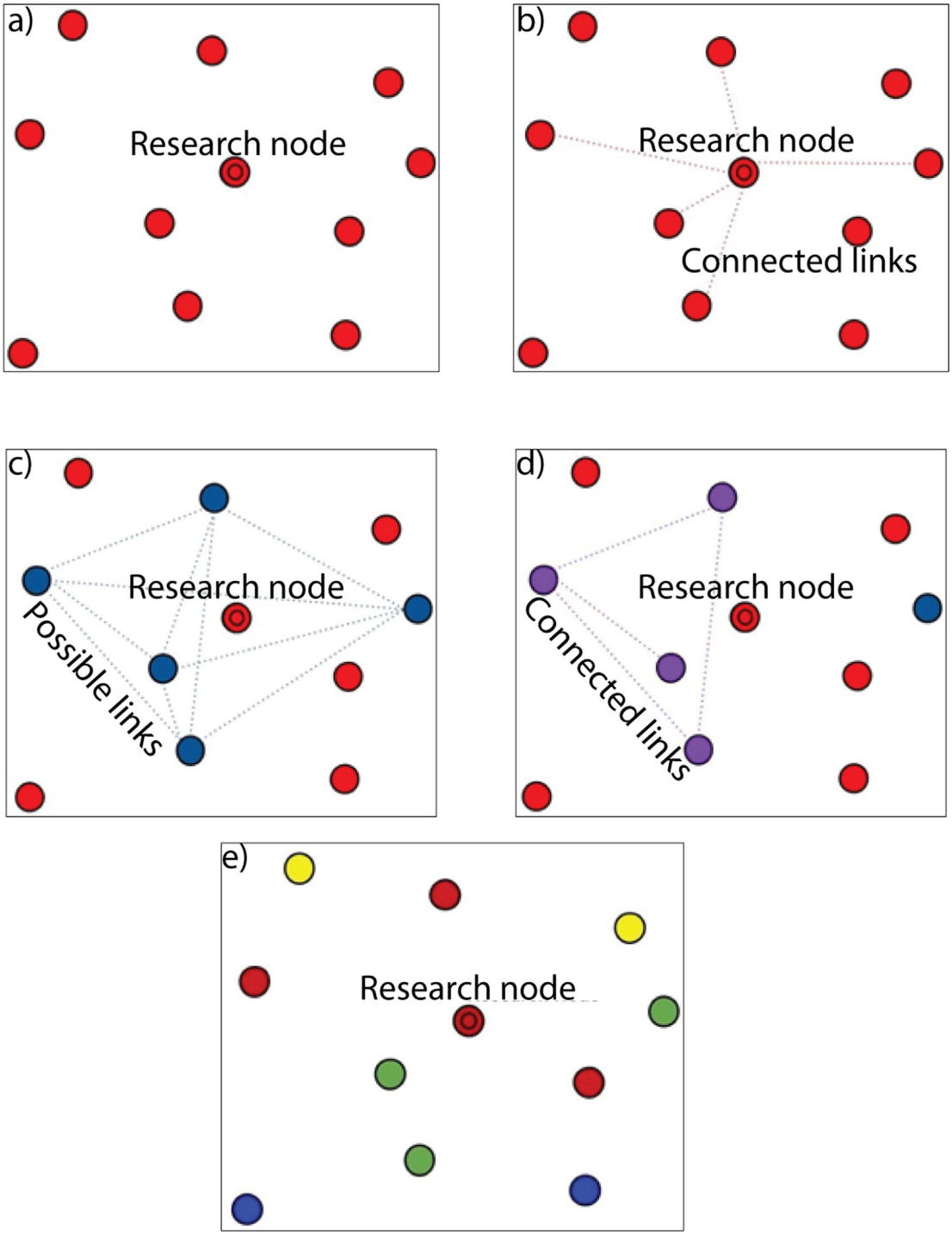

Fig. 1. Three variants of IDW, (a) standard IDW with nearest neighbors (IDW_NN); (b) to (d) shows steps in estimating linked neighbors for the model, IDW_LN; and (e) IDW with clustered neighbors (IDW_CN).

geographically closer, rainfall data recorded at seemingly neighboring stations can vary significantly due to the variability in topography; see, for instance, Berndtsson (1988) and Li et al. (2014) for additional details in relation to spatio-temporal variability. Spatial patterns are consistently influenced by topography and wind direction, especially in mountainous areas (Barros and Lettenmaier, 1993; Barry, 1992). Considering these limitations, a better understanding of spatial connections between rain gauges and the effect of including such information in IDW method need to be explored.

In recent years, the theory of networks has been widely applied for studying the spatial and temporal evolution of a wide range of complex systems and associated phenomena (Barabási and Albert, 1999; Girvan and Newman, 2002; Konapala and Mishra, 2017; Li et al., 2010; Milo et al., 2002). The application of network theory in hydrology and water resources is relatively new with increasing number of publications on the topics of connections in rainfall, stream flow, river networks, and virtual water trade networks; see Sivakumar (2015) for a general account of this topic.

As for rainfall, Malik et al. (2012) investigated the spatial and temporal characteristics of extreme (summer) monsoon rainfall in South Asia. Boers et al. (2013) used networks based concepts to investigate the South American Monsoon System (SAMS) spatial

characteristics of extreme rainfall synchronicity by analyzing gridded daily rainfall data; see also Boers et al. (2015) for subsequent complex networks-based studies on South American rainfall. In order to examine the annual dynamics of precipitation around the world, Scarsoglio et al. (2013) analyzed a 70 year long (January 1941–December 2010) gridded precipitation dataset. In the course of the analysis of monthly rainfall data for a period of 68 years at 230 rain gauge stations across Australia, Sivakumar and Woldemeskel (2015) employed the concepts of clustering coefficient and degree distribution for the examination of spatial connections in rainfall. The clustering coefficient is a measure of local density and represents the tendency of a network to cluster (Watts and Strogatz, 1998). Jha et al. (2015) attempted to provide a hydrological explanation for the results of the complex network-based methods for rainfall. The clustering coefficient method was applied to two different rain gauge networks in Australia (57 stations in Western Australian stations and 45 stations in the Sydney region) and the results were interpreted as topographical properties of rain gauge stations (latitude, longitude and elevation) and rainfall data characteristics (mean, standard deviation and variation coefficient) (Jha and Sivakumar, 2017). Naufan et al. (2018) studied the spatial connections in rainfall data from a regional climate model, in the context of climate change.

With the encouraging results detailed by these preliminary studies, the present study aims to apply network theory to take into account the spatial connections across a rain gauge network into the IDW approach to reconstruct the rainfall data at a desired location. We use the concepts of networks to analyze the significance of spatial connections and spatial correlation in the rainfall network. We propose three variants of Inverse distance weighing; i.e., IDW_NN (nearest neighbors) model, IDW_LN (linked neighbors) model and IDW_CN (clustered neighbors) model to study the significance of spatial and temporal connection over spatial correlation. For implementation, we consider the daily rain gauge data of 430 rain gauge stations, located in the Murray Darling Basin. To evaluate the performance of the proposed models, we use the one-station-leave-out cross validation approach. We also study the effect of location and elevation of the rain gauges on the performance of the proposed models.

2. Methodology

2.1. IDW with nearest neighbors (IDW_NN)

We use the IDW method in its standard form to estimate the rainfall at a desired location using weighted average of rainfall at 'n' nearest neighbors as shown in Fig. 1(a) (referred as IDW_NN hereafter). The following formula is used:

$$P_i = \frac{\sum_{x=1}^{n}\left[\frac{1}{d_x{}^t} \times P_x\right]}{\sum_{x=1}^{n}\frac{1}{d_x{}^t}} \tag{1}$$

where P_i is the estimate of rainfall at the desired location i; P_x represents rainfall at a neighboring location x; $d_x$is the distance from the location x to the location i where rainfall is being estimated; and n is the number of neighborsand t is a positive real number, called the power parameter. For the sake of simplicity, $t = 2$ is considered in further IDW calculation. Eq. (1) estimates the weighted average of the rainfall recorded at the nearest 'n' rain gauges. To select a fixed value of the number of neighbors in the application of IDW_NN, a sensitivity analysis on the number of neighbors is performed. For the sensitivity analysis, the nearest neighbors are varied from 1 to 15 in the calculation of IDW_NN and the error statistics associated with the reconstruction of the data are evaluated (more details about the sensitivity analysis are provided in Section 4.1).

2.2. IDW with linked neighbors (IDW_LN)

In the case of IDW_NN, the neighbors are generally determined based on the geographical proximity only, as is done in the present study as well. Here, we introduce the concept of network to identify the neighbors. A network is a set of points connected together by a set of lines, as shown in Fig. 1(b). The points are called nodes or vertices, and the lines are referred as links or edges. Mathematically, a network can be represented as G = [P, E], where P is a set of N nodes ($P1$, $P2, \cdots, PN$) and E is a set of M links. In the present context, rain gauges can be considered as nodes of the network and the connections among them will be the links. There are various ways and measures to study the network properties, such as clustering coefficient, degree centrality, and degree distribution; see Newman (2012) and Estrada (2012) for details. The clustering coefficient is used in this study to investigate the spatial connections in the rain gauge network. The clustering coefficient (CC) is a measure of the local density of a network and quantifies the network's tendency to cluster.

To find the clustering coefficient, the first step is to assign a correlation threshold CT to identify the neighbors of research station i, i.e. links that have correlations exceeding CT. We refer the number of such neighbors as k_i. For instance, suppose there are 10 rain gauges surrounding the research node. Out of 10, there are only 5 rain gauges at which the correlation of rainfall at the research node 'i' exceeds the preselected CT value as shown in Fig. 1(b). Then there would be $\frac{k_i(k_i-1)}{2}$possible links among 'k_i' neighbors of research station 'i' (blue lines in Fig. 1(c)) (see also Sivakumar and Woldemeskel (2014) for additional details).

The second step in the estimation of clustering coefficient is to find the possible links between k_i nodes which also exceed beyond CT. Let E_i be the number of links among neighboring stations with correlations exceeding CT (four links shown in violet lines in Fig. 1(d). We refer the corresponding nodes as linked neighbors of the research station. We propose a variant of IDW_NN in which we consider only linked neighbors in the IDW approach, which we refer as IDW_LN. Furthermore, to study the sensitivity of CT on the reconstruction of data using IDW_LN, the CT values are varied from 0.3 to 0.9 (more details about the sensitivity analysis are provided in Section 4.3).

2.3. IDW with clustered neighbors (IDW_CN)

The third step in the calculation of CC at the research station is to use formula:

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

where E_i is the number of links that actually exist between these k_i stations and $\frac{k_i(k_i-1)}{2}$ are the total number of possible links between these k_i stations. The procedure to find the value of E_i and k_i is repeated for each and every node (research gauge station) in the network to obtain the clustering coefficient at the corresponding node. Once we get the CC value for all the nodes in the network, different Clustering Coefficient Ranges (CCR) can be defined to classify the nodes into various groups (for example: four different clusters shown in red, green, yellow and blue in Fig. 1(e)). Now, as another variant of IDW_NN, we can consider only those nodes as neighbors which belong to the same CC range as the research node in applying IDW; we refer this as IDW_CN hereafter. Further, to study the sensitivity of CCR on the reconstruction of data using IDW_CN, three CCR are fixed, i.e. 0.3 to 0.6, 0.4 to 0.7 and 0.5 to 0.8.

2.4. Verification of the interpolated rainfall

Once the rainfall has been reconstructed at the research node using one of the three models (IDW_NN, IDW_LN, IDW_CN), we estimate the

root-mean-square error (RMSE) and the bias percentage (BP) at that node as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P(x_i) - P^*(x_i))^2} \quad (3)$$

$$BP = \sum_{i=1}^{n}\frac{(P(x_i) - P^*(x_i)) \times 100}{P^*(x_i)} \quad (4)$$

where $P(x_i)$ is the reconstructed rainfall and $P^*(x_i)$ is the observed rainfall at station x_i. Lower value of RMSE and BP associated with the reconstruction of data implies the better performance of the model used in the reconstruction of the data. The RMSE measures the average magnitude of the error; it is the representation of the data around the line of best fit. The RMSE does not necessarily increase with the variance of the errors but increases with the variance of the frequency distribution of error magnitudes. However, the BP measures the average tendency of the interpolated values to be larger or smaller than their observed ones.

2.5. Experimental setup

We apply the one-station-leave-out cross validation approach to assess the performance of the IDW_NN, IDW_LN, and IDW_CN models. One-station-leave-out cross validation approach leaves one station out of the training data, i.e. if there are *n* stations in the original sample then, *n-1* stations are used to train the model and the selected station is used as the validation station. This is repeated for all combinations in which the original sample can be separated this way, and then the error is averaged for all trials, to give overall effectiveness. It is used in determining the hyper parameters of a model, in the sense that it identifies which parameters will result in the lowest test error. As described in sections 2.1 to 2.3, in applying three different IDWs, the number of neighbors has to be determined. Also, in the case of IDW_LN and IDW_CN, a suitable correlation threshold needs to be used. We perform our analysis in the following three steps to systematically find out (i) the sensitivity of number of neighbours, correlation threshold, and clustering coefficient range on the reconstructed data, (ii) the relationship of the error statistics associated with the reconstruction of rainfall data with the elevation of the rain gauge stations, and (iii) the spatial distribution of errors in the reconstructed rainfall data.

3. Study area and dataset

We consider rainfall data from the largest river basin in Australia, namely the Murray-Darling Basin (MDB) (Fig. 2). The MDB contains the catchment of the Murray River (length 2508 km) and the Darling River (length 1472 km) covering a large portion of south-east area (approximately 106 km^2) of the Australian continent. The basin extends from -138.81^o to -152.42^o longitudes and -37.51^o to -25.32^o latitude. Most regions of the basin are flat and low-lying but the coastal areas of the basin contain mountains (the Great Dividing Range). The climate in the MDB is mostly semi-arid causing variability in rainfall at different temporal scales that affect Australian agriculture. Daily rainfall data from 430 rain gauges for the period of January 1985 to December 2014 were collected from the Bureau of Meteorology (BoM) in Australia. Table 1 provides a summary statistic of the rainfall data from the 430 rain gauges at the daily scale.

4. Results

To evaluate the relative improvement, if any, in the reconstructed rainfall data using the concept of networks, we devise two sets of experiments: compare results of IDW_LN with IDW_NN, and that of IDW_CN with IDW_NN. The results from IDW_NN are used as reference. The number of neighbors in IDW_LN and IDW_CN models is expected to vary by changing CT, clustering coefficient range (CCR) and other relevant parameters. Further, in special cases when, for instance, CT is very high; a research node may not have any neighbor at all. To avoid such scenarios in the analysis, we introduce the concept of a valid station, the interpretation of which is slightly different in case of IDW_LN and IDW_CN models.

4.1. Sensitivity of the number of neighbors in IDW_NN model

To perform the sensitivity analysis, we vary the number of neighbors from 1 to 15 and use leave-one-station-out cross validation at each rain gauge station. The RMSE and BP values presented in Fig. 3a and 3b, respectively, show the error values averaged over 430 rain gauge stations. It is clear from the plots that both average RMSE and average BP decrease (except average BP for 6 neighbors where a slight increase is seen) as the number of neighbors increases from 1 to 15. The magnitude of average RMSE values varies from 3.11 to 3.86, while the average BP values vary from 16 to 18.5. Low error statistics demonstrate that the IDW_NN model is able to reconstruct the rainfall data accurately over a major part of the basin. Since the ranges of RMSE and BP are low, it can also be concluded that the variation in the error values with the increase in the number of neighbors is not significant. The purpose of the sensitivity analysis was to select a fixed number of neighbors which can be used in later analysis using IDW_LN and IDW_CN. Fig. 3 shows that after number of neighbors equals to five, there is no significant change in the average error values; hence, we decide to use five neighbors in further analysis. The average RMSE with 5 neighbors in IDW_NN is 3.19 and the associated average BP is 16.166, as shown in Fig. 3. In future analysis, we will use these error statistics (RMSE = 3.19, and BP = 16.17) to compare the performance of the IDW_LN and IDW_CN models.

4.2. Concept of a valid station

A valid station is a rain gauge station for which a specific model (IDW_LN or IDW_CN) will have sufficient number of neighbors for the reconstruction of rainfall. For the IDW_NN model, all rain gauges are valid stations because the selection of neighbors is independent of a certain CT or CCR. In the case of IDW_LN model, we define a valid station only when it has at least five neighbors (as obtained in the previous section) for a given CT. Similarly, for the IDW_CN model, a station with at least five neighbors for a given CT and CCR is considered a valid station.

4.3. Sensitivity of the CT in IDW_LN model

After fixing the number of neighborhood to five, the next task is to fix the correlation threshold used in the IDW_LN model. The sensitivity analysis of CT is performed by changing its values from 0.3 to 0.9 in the IDW_LN model. It is expected that with the increase in CT, the number of valid stations would go down. Table 2 shows that out of 430 rain gauge stations, 427 are identified as valid stations by applying a very low threshold on correlation as 0.3. With the increase in CT up to 0.6, there are minor changes in the number of valid stations; however, there was a major drop in the number of valid stations for CT = 0.7 and higher. Both the IDW_NN and IDW_LN models were applied at the same set of valid stations to estimate the error statistics, RMSE and BP. For instance, 422 valid stations exist by using CT = 0.4; then for each of these 422 stations, we use 5 geographically nearby neighbors (as decided in Section 4.2) to apply IDW_NN; on the other hand, we use 5 linked neighbors in the calculation of IDW in case of IDW_LN. The average values of RMSE and BP correspond to errors averaged over all the valid stations (e.g., 422 in case of CT = 0.4). Table 2 shows that the average RMSE decreases as CT increases which is valid for both IDW_NN and IDW_LN models. By comparing the third and fourth columns in Table 2, it can be concluded that the value of average RMSE is relatively higher in the case of reconstructed rainfall using IDW_LN.

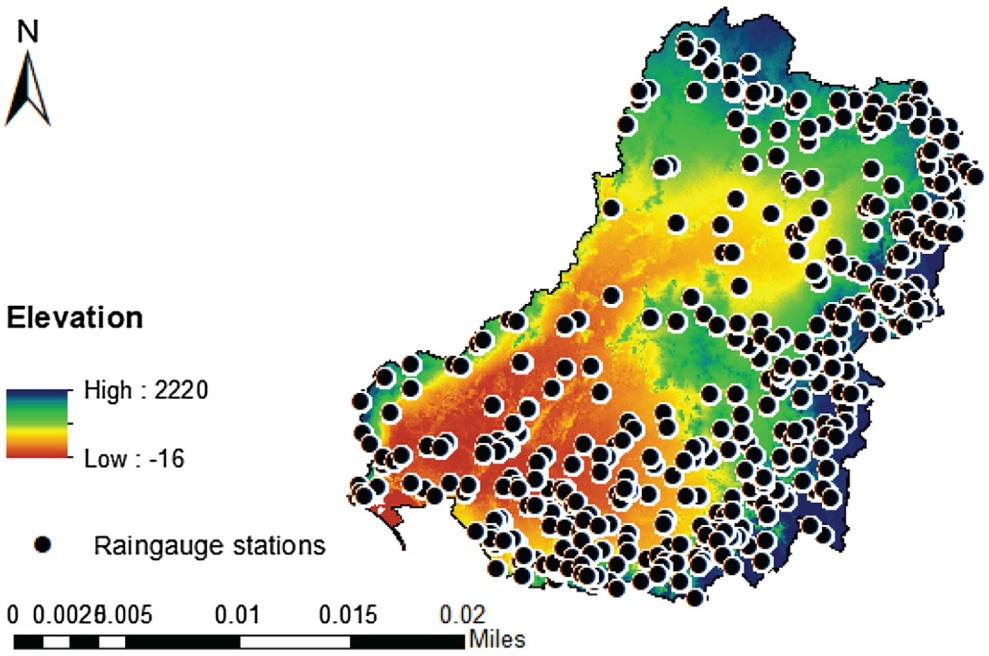

Fig. 2. The Murray-Darling Basin with the locations of 430 rain gauges.

Table 1
Summary statistics of rainfall at 430 gauging stations at daily temporal scale.

Statistics	Value
Length of data	10957 days
Range of mean (mm)	0.21–3.79
Range of Standard Deviation (mm)	1.84–10.32
Range of maximum rainfall (mm)	53–482

Similar observation is also made in the case of average BP (as shown in Table2). Since these results are averaged values of errors over all the valid stations, we decide to explore what percentage of valid stations actually shows less RMSE and/or BP using the IDW_LN model. From Table 2, it can be seen that at about 28.5 to 31.8% of valid stations, RMSE in reconstructing the data using the IDW_LN model is lower than that from the IDW_NN model. Similarly, at about 28.5 to 50.3% of valid stations, BP values are found to be less when the IDW_LN model is used in reconstructing the rainfall data. The average RMSE and BP values associated with both IDW_NN and IDW_LN are comparable.

The next obvious question is to determine at each of the valid stations, whether the RMSE and BP values are higher from the IDW_NN model or from the IDW_LN. Fig. 4 shows the scatter diagram between the errors associated with each valid station from both of these IDW models. In Fig. 4, if the points fall on the identity line, it means that at a valid station, the errors in the reconstructed rainfall data from the IDW_NN and IDW_LN models are equal. If a point falls above or below the identity line, it will be interpreted as the associated error from the IDW_LN model is higher or lower, respectively. For the sake of simplicity in presentation, plots corresponding to only 3 CT values, i.e., 0.4, 0.6 and 0.8 are shown in Fig. 4. As seen, the cloud of points decreases as CT increases because the number of valid stations decreased

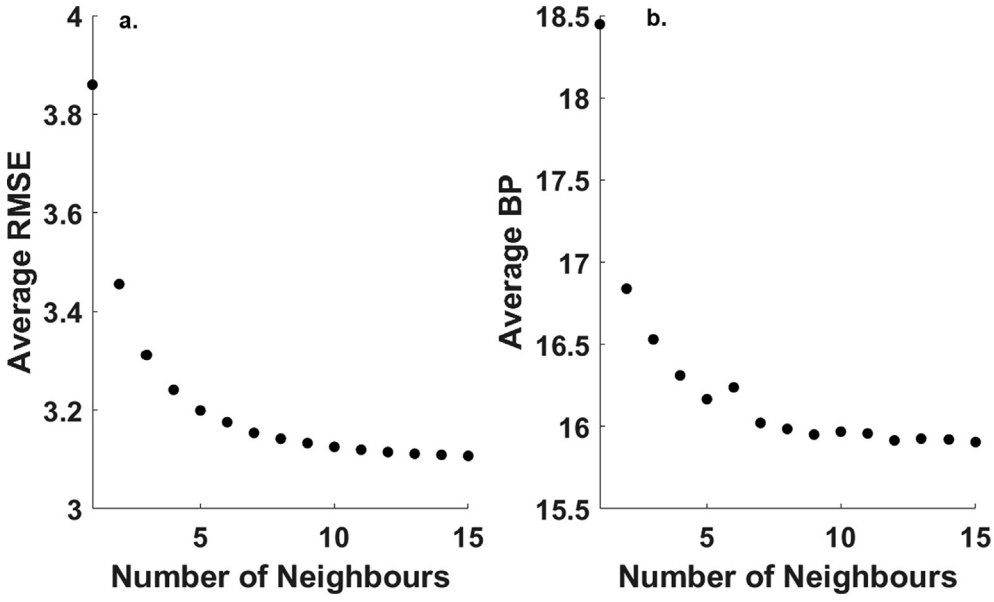

Fig. 3. Plots of average error associated with IDW_NN against the number of neighbors in the inverse distance weighing application.

Table 2
Summary statistics of errors (RMSE and BP) associated with the IDW_LN models.

Error statistic	Number of valid stations	Average RMSE		% valid stations showing less error with IDW_LN	Average BP		% valid stations showing less error with LN
Model Type		IDW_NN	IDW_LN		IDW_NN	IDW_LN	
CT							
0.300	427	3.185	3.332	29.977	12.86	14.072	44.965
0.400	422	3.172	3.319	29.858	11.408	12.796	45.972
0.500	398	3.131	3.282	29.397	9.745	10.886	48.744
0.600	304	2.954	3.112	31.579	8.970	9.716	50.329
0.700	132	2.629	2.737	31.818	9.395	10.599	46.212
0.800	7	2.395	2.431	28.571	8.596	10.399	28.571
0.900	0	NaN	NaN	NaN	NaN	NaN	NaN

significantly from 422 to 7 (see Table 2).

In Fig. 4(a) to (c), as the points are close to the identity line, the difference in the associated RMSE with IDW_NN and IDW_LN is quite low. On the other hand, the plots of the BP error, indicate in Fig. 4(d) to (f) shows a much wider scatter than the RMSE error, mainly because the percentage bias is relative and RMSE is absolute error.

4.4. Variation of errors in IDW_LN with location of valid stations

Since there is a large topographical variation in MDB (see Fig. 2), we investigate into whether there is a relationship between the errors in reconstructing the rainfall data and the elevation of the rain gauge station when the IDW_NN and IDW_LN models are used. In this direction, the first task is to observe the location of the valid stations located in the MDB where the IDW_LN model performed better than IDW_NN model. The aim here is to examine whether by introducing the concept of network, the reconstruction model is able to take into account the topographical variations better or not.

Fig. 5 shows the performance of the IDW_LN model compared to the IDW_NN model as a function of topography of the rain gauge stations. Fig. 5(a) shows the location of the valid stations with lower RMSE and BP using the IDW_LN model compared to errors associated with the IDW_NN model for CT = 0.3 to 0.6 (Since the number of valid stations decreased significantly after CT = 0.6, we choose to present the locations of the valid stations in the MDB only corresponding to four CT values, i.e. 0.3 to 0.6). From both the rows of Fig. 5(a), it is evident that the number of valid stations decreases with an increase in the CT value from 0.3 to 0.6. The north-west part of the MDB has the least number of valid stations for IDW_LN (empty square in Fig. 5(a)), which implies that it is the least-connected region of the basin based on the correlation of rain gauge data. Similarly, the second row of Fig. 5(a) shows that the valid stations corresponding to lower BP are more or less equally distributed in the MDB.

Fig. 5(b) shows the performance of the IDW_LN model as a function of different elevation bands of the rain gauges. The whole rain gauge network is divided into six elevation bands, i.e. below 200 m, 200 to 400 m, 400 to 600 m, 600 to 800 m, 800 to 1000 m, and above 1000 m. For each elevation band, Fig. 5(b) presents the bar plots of the total number of rain gauges associated with the band, total number of IDW_LN valid stations associated with the band, total number of valid stations with lower RMSE from IDW_LN as compared to IDW_NN and total number of valid stations with lower BP from IDW_LN as compared to IDW_NN. For simplicity, only CT = 0.6 is considered for presentation in Fig. 5(b). It can be concluded, from Fig. 5(b), that the rain gauge stations where the IDW_LN model showed lower RMSE than the IDW_NN model is equal to nearly 30 percent of the valid stations in that particular elevation band. This implies that the performance of the model is independent of the elevation of the rain gauge stations. The rain gauge stations which yielded lower BP from IDW_LN than the IDW_NN model is equal to nearly 50 percent of the valid stations in all the elevation bands (see Fig. 5(b)). Thus, it is not straightforward to derive any definite conclusion about the effect of topography on the

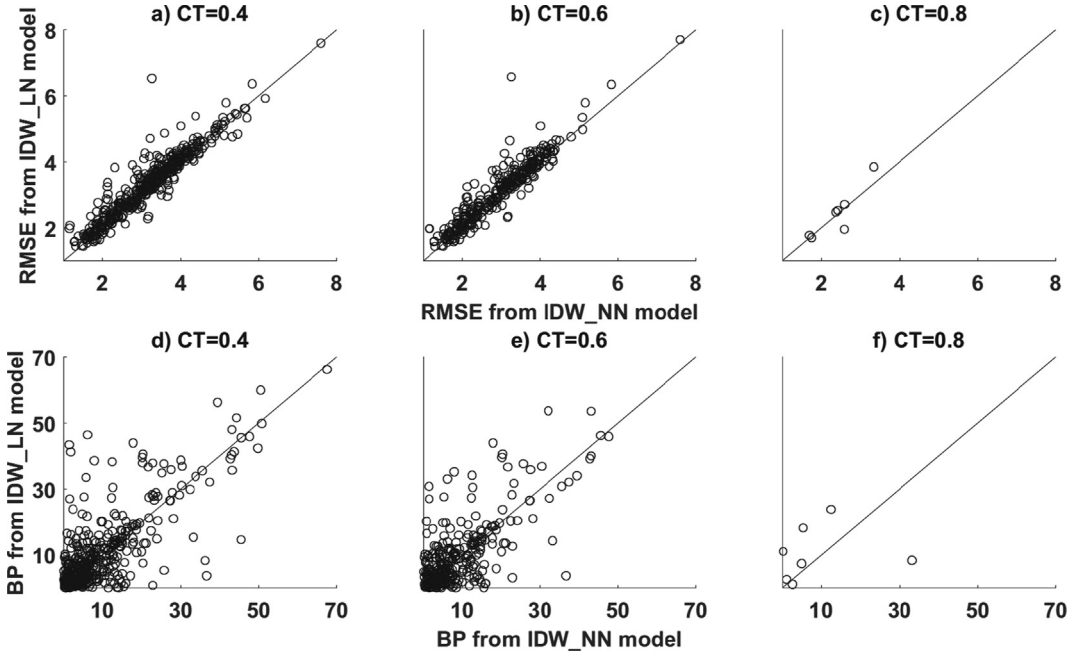

Fig. 4. Comparison of errors between IDW_NN model and IDW_LN model for all valid stations.

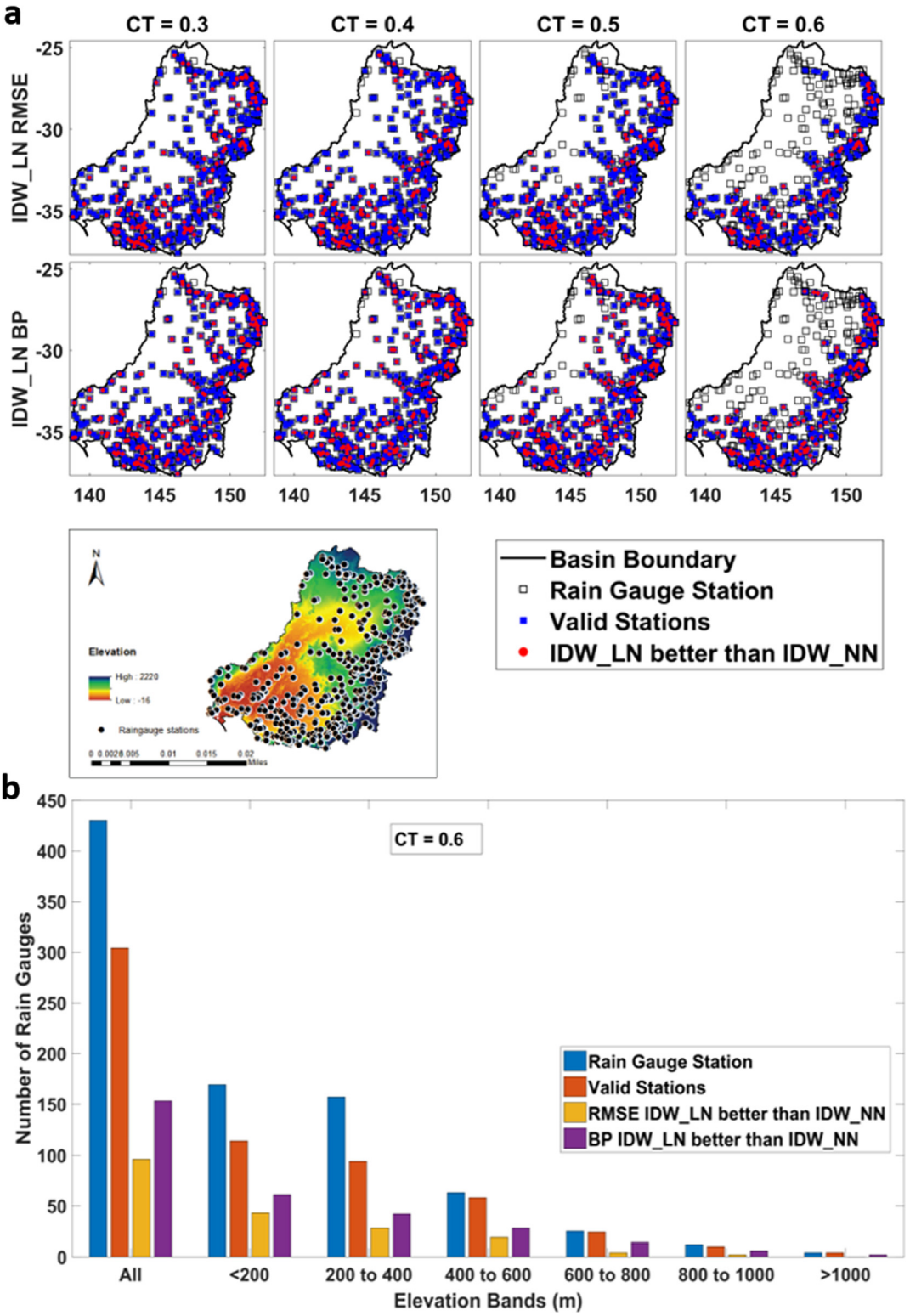

Fig. 5. Performance of the IDW_LN model with (a) the locations of the valid stations; and (b) the elevation of the valid stations.

performance of IDW_LN and IDW_NN models in the reconstruction of the rainfall using either the IDW_LN model or the IDW_NN model (see Fig. 5(a) and (b)).

4.5. Variation of errors in the IDW_LN model with elevations of valid stations

Next, we explore the relationship between the elevations of the valid stations and the magnitude of the associated error estimates from the IDW_LN model in reconstructing the rainfall data. In Fig. 6, the RMSE and BP values associated with each of the valid stations are plotted against its elevation for CT = 0.4, 0.6 and 0.8 (For simplicity, we present the results for only these three selected CT values). As it is clear from the scatter of values in Fig. 6(a) to 6(c), the valid stations with low RMSE obtained from the IDW_LN model are situated at low elevations (mostly below 500 m with RMSE value up to 3.5, see

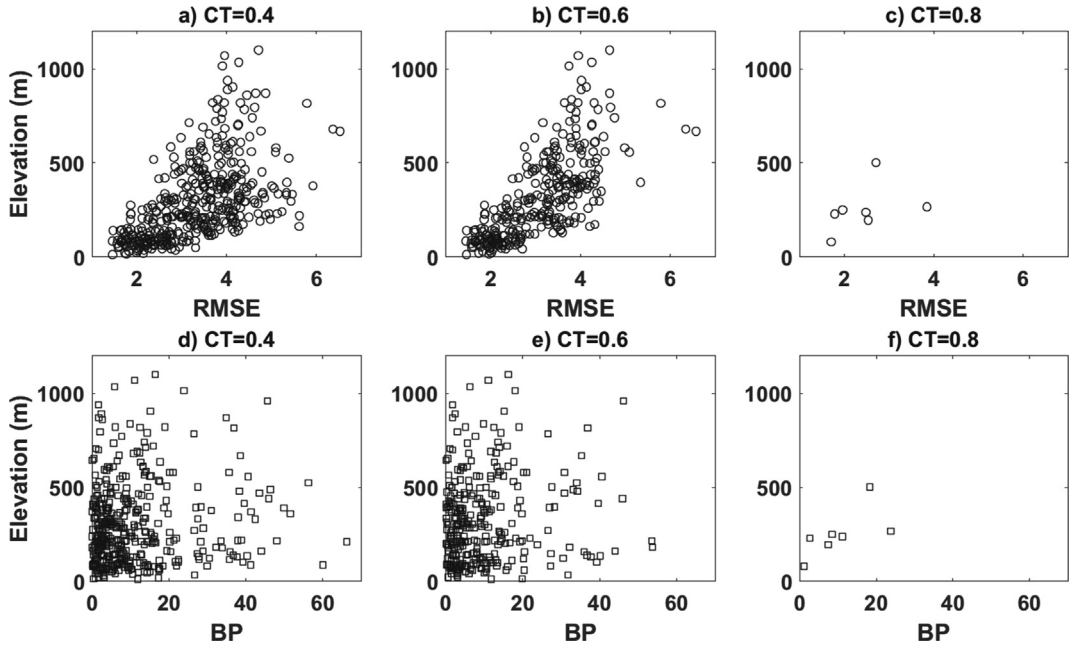

Fig. 6. Plots of elevation and error (RMSE and BP) associated with all valid stations for the IDW_LN model.

Table 3
Summary statistics of RMSE associated with the IDW_CN mode.

Error statistic	[CCR = 0.3 to 0.6]				[CCR = 0.4 to 0.7]				[CCR = 0.5 to 0.8]			
Model type	Number of valid stations	Average RMSE		% Valid stations showing less error with IDW_CN	Number of valid stations	Average RMSE		% Valid stations showing less error with IDW_CN	Number of valid stations	Average RMSE		% Valid stations showing less error with IDW_CN
		IDW_NN	IDW_CN			IDW_NN	IDW_CN			IDW_NN	IDW_CN	
CT												
0.300	421	3.189	3.344	30.166	302	3.126	3.346	28.808	146	3.267	3.538	26.712
0.400	394	3.162	3.335	28.934	243	3.287	3.573	23.457	98	3.711	4.207	19.388
0.500	331	3.239	3.470	25.378	184	3.428	3.788	15.217	90	3.687	4.294	12.222
0.600	259	3.182	3.495	15.058	161	3.366	3.868	9.317	94	3.602	4.289	7.447
0.700	177	2.970	3.373	10.734	120	3.139	3.697	5.000	97	3.218	3.928	1.031
0.800	43	2.712	3.485	9.302	28	2.800	3.785	7.143	28	2.800	3.785	7.143
0.900	0	NaN	NaN	NaN	0	NaN	NaN	NaN	0	NaN	NaN	NaN

Table 4
Summary statistics of BP associated with the IDW_CN model.

Error statistic	[CCR = 0.3 to 0.6]				[CCR = 0.4 to 0.7]				[CCR = 0.5 to 0.8]			
Model type	Number of valid Stations	Average BP		% Valid stations showing less error with IDW_CN	Number of valid stations	Average BP		% Valid stations showing less error with IDW_CN	Number of valid stations	Average BP		% Valid stations showing less error with IDW_CN
		IDW_NN	IDW_CN			IDW_NN	IDW_CN			IDW_NN	IDW_CN	
CT												
0.300	421	16.411	17.574	45.131	302	20.222	22.387	44.702	146	31.797	30.968	46.575
0.400	394	11.808	13.015	46.447	243	14.887	15.992	48.184	98	21.164	23.663	43.878
0.500	331	11.541	13.431	45.921	184	14.323	15.788	47.826	90	19.919	22.260	41.111
0.600	259	9.401	11.329	45.560	161	10.840	14.024	45.963	94	12.273	16.032	39.362
0.700	177	8.259	10.6433	41.808	120	8.686	11.223	41.667	97	8.208	11.271	38.144
0.800	43	10.115	16.780	37.209	28	10.729	21.157	28.571	28	10.729	21.157	28.571
0.900	0	NaN	NaN	NaN	0	NaN	NaN	NaN	0	NaN	NaN	NaN

Fig. 6(a)). The rain gauge stations situated at higher elevations in the MDB basin (above 800 m) show higher RMSE (values above 4). The BP values associated with the IDW_LN model are relatively low (mostly below 30) at the high elevation areas of the basin (above 500 m) and reach to as high as 60 for low elevation areas (below 300 m) of the basin, which is mainly because the sample size of the lower elevation bands is larger (Fig. 5(b)) as compared to that at the higher elevation bands.

4.6. Sensitivity of CT and CCR in the IDW_CN model

After studying the performance of the IDW_LN model, now we analyze the error statistics associated with the IDW_CN model. The sensitivity analysis of the correlation threshold (CT) and the clustering

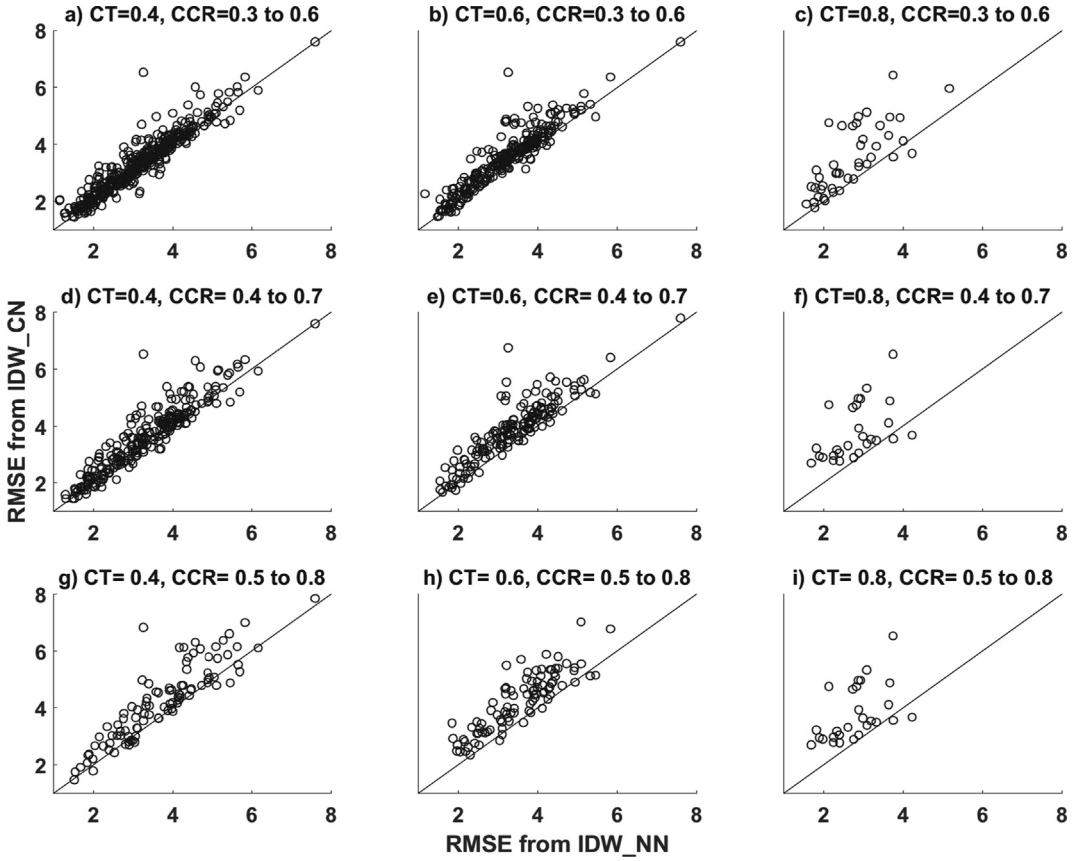

Fig. 7. Plots of IDW_CN model RMSE and the corresponding IDW_NN model errors for all valid stations.

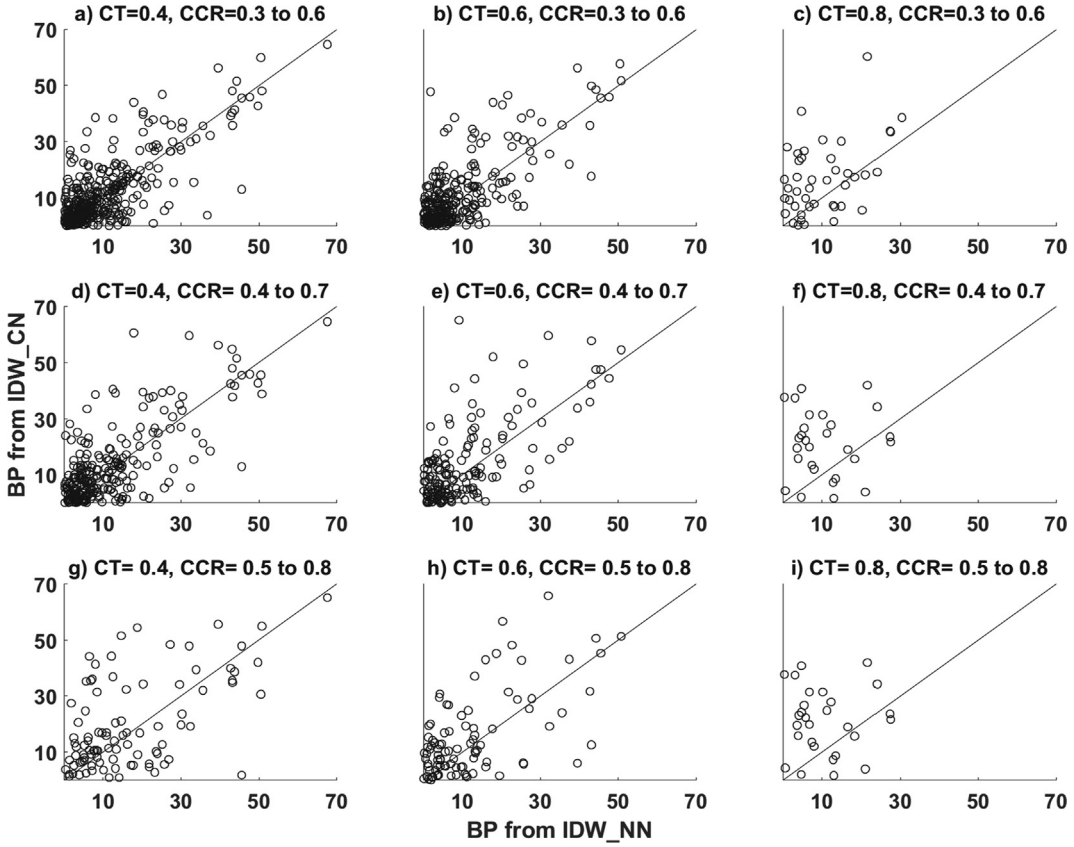

Fig. 8. Plots of IDW_CN model BP and the corresponding IDW_NN model errors for all valid stations.

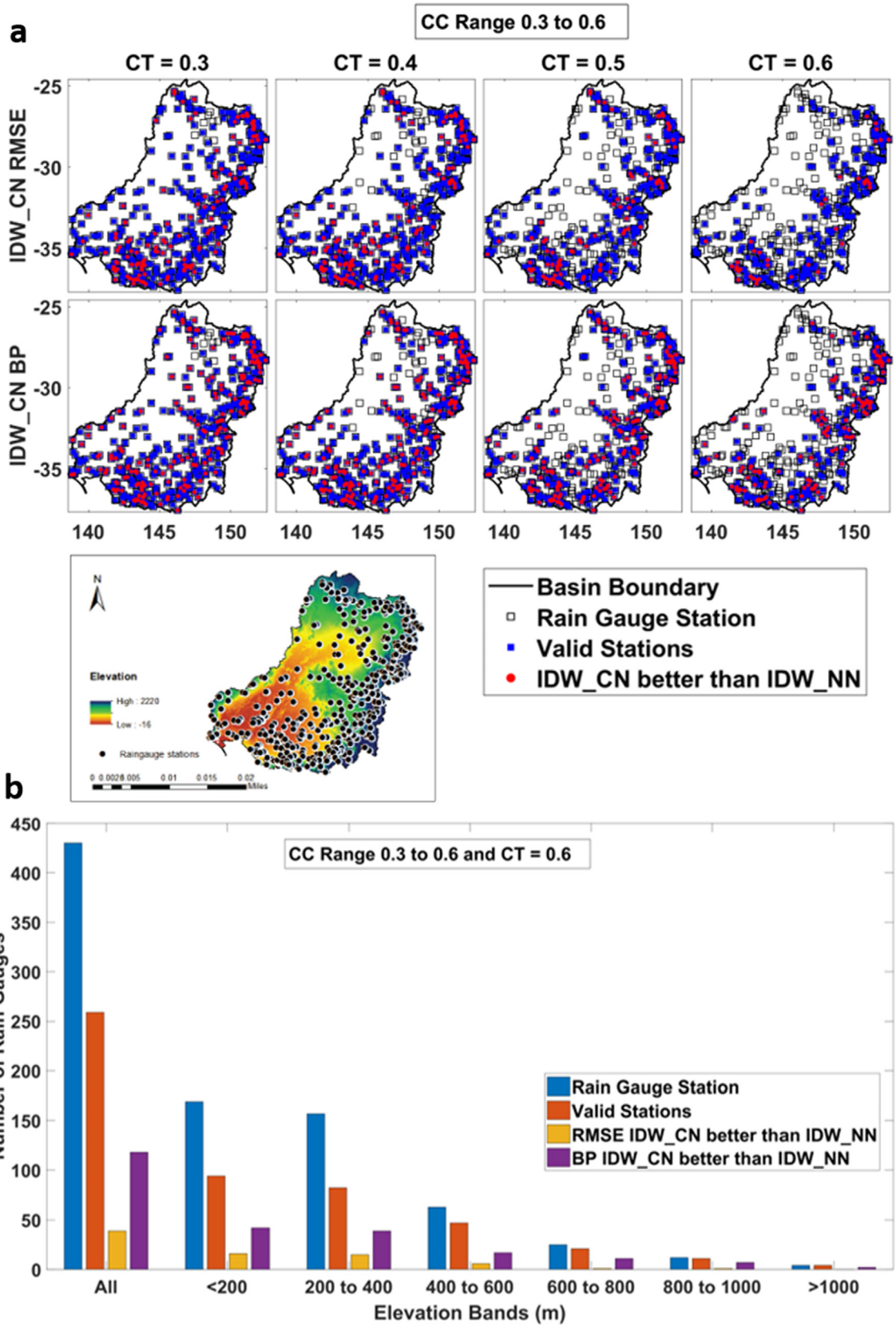

Fig. 9. Performance of the IDW_CN model with (a) the locations of valid stations; and (b) elevations of valid stations.

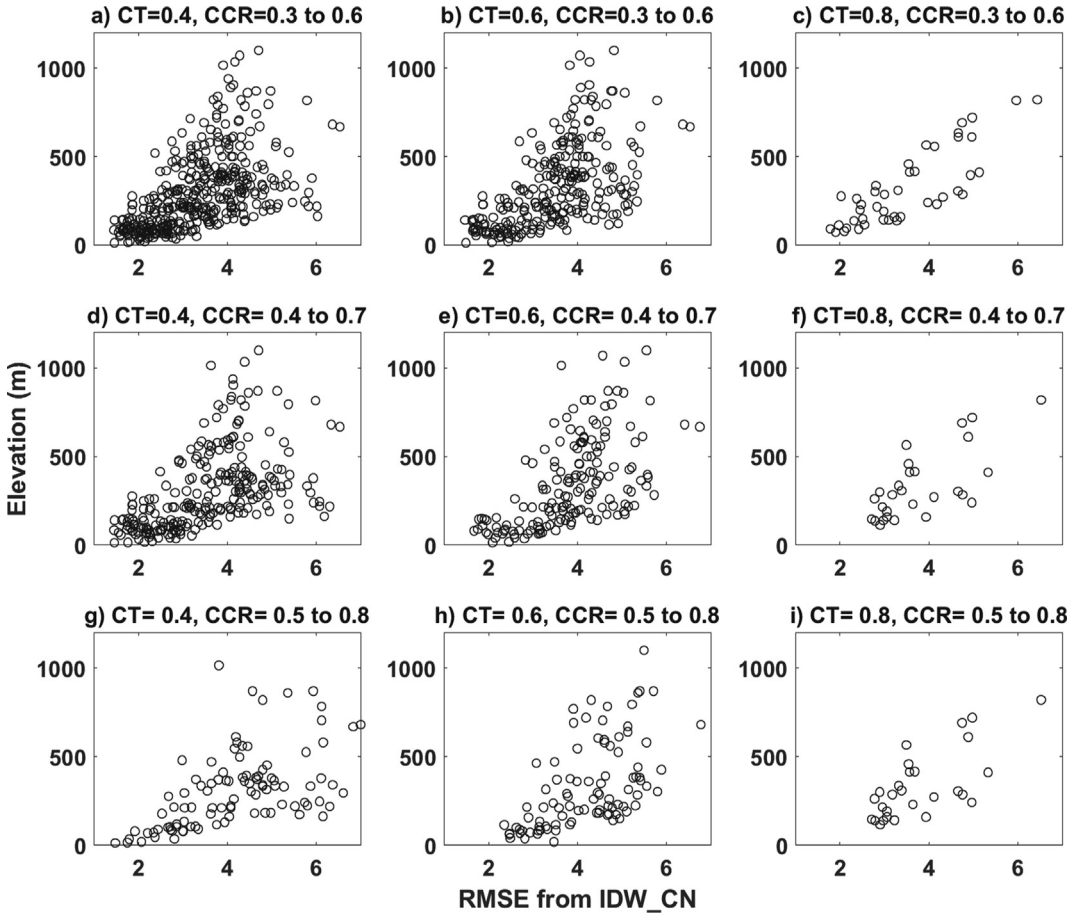

Fig. 10. Plots of elevation and RMSE associated with all valid stations for IDW_CN model.

coefficient range (CCR) are performed by changing the CT values from 0.3 to 0.9 for three different CC ranges: 0.3 to 0.6, 0.4 to 0.7 and 0.5 to 0.8 respectively in the IDW_CN model. Tables 3 and 4 show that out of the 430 rain gauge stations, for CCR equal to 0.3 to 0.6, the number of valid stations ranges from 421 to 0 for different CT values (0.3 to 0.9). As the CCR increases to 0.5 to 0.8, the selected valid station range decreases i.e. from 146 to 0. These observations indicate that for higher CCR, the number of valid stations decreases significantly because the clusters become more specific for higher CT and CCR.

In terms of the magnitude of errors, Table 3 shows that the average RMSE increases as the CCR increases for each CT value (0.3 to 0.9), which is valid for both the IDW_NN and IDW_CN models. By comparing the average RMSE associated with the IDW_NN and IDW_CN models in each CCR, it can be concluded that the value of the average RMSE is relatively higher in the case of reconstructed rainfall using the IDW_CN model than that using the IDW_NN model for all CCR values. Similar observation is made also in the case of the average BP (as shown in Table 4).

Further, to compare the RMSE and BP from the IDW_CN model or from the IDW_NN model at each of the valid stations, plots of RMSE and BP associated with the IDW_CN model with the corresponding IDW_NN model errors for all the valid stations are generated. Figs. 7 and 8 show that the scatter between the errors associated with each valid station from both of these IDW models for three CT values (0.2, 0.4, 0.6) and 3 CCR values (0.3 to 0.6, 0.4 to 0.7, 0.5 to 0.8). In both figures, the cloud of points decreases as the CT and CCR increases because the number of valid stations decreases significantly at higher CT and CCR (see Tables 3 and 4). As shown in Fig. 7, the RMSE scatter plots shift above the identity line as CT and CCR increases, which indicates that the error associated with a valid station increases as CT and CCR increases in the IDW_CN model. In Fig. 8, the points are scattered almost evenly around the identity line, hence the associated BP with the IDW_CN model and the IDW_NN model differs from low to high value.

Tables 3 and 4 also list the percentages of valid stations showing less RMSE and/or BP, respectively, using the IDW_CN model, for three different CCR values. The percentage of valid stations with lower RMSE and BP associated with the IDW_CN model than that from the IDW_NN decreases as the CT and CCR increase. Hence, the IDW_CN model works fairly well at lower CT and CCR values when compared to their higher values.

4.7. Variation of errors in the IDW_CN with locations of valid stations

To investigate the relationship of errors in reconstructing rainfall data using the IDW_NN and IDW_CN models with the topography of the rain gauge station, we identify the valid station locations in the MDB where the IDW_CN model performed better than the IDW_NN model. The motivation here is to examine whether by introducing the concept of clusters, the reconstruction model is able to take into account the topographical variations better or not. Fig. 9(a) shows the performance of the IDW_CN model compared to that of the IDW_NN model as a function of the locations of the rain gauges for CCR = 0.3 to 0.6. As the number of valid stations for the CCR values 0.4 to 0.7 and 0.5 to 0.8 are significantly low (Tables 3 and 4), the locations of valid stations for these two CCR values are not plotted, for the sack of simplicity. The locations of valid stations associated with lower RMSE and BP at different CT and CCR do not show any particular trend and are randomly located in the case of the IDW_CN model. This particular observation may be due to the specific nature of clusters associated with the IDW_CN model.

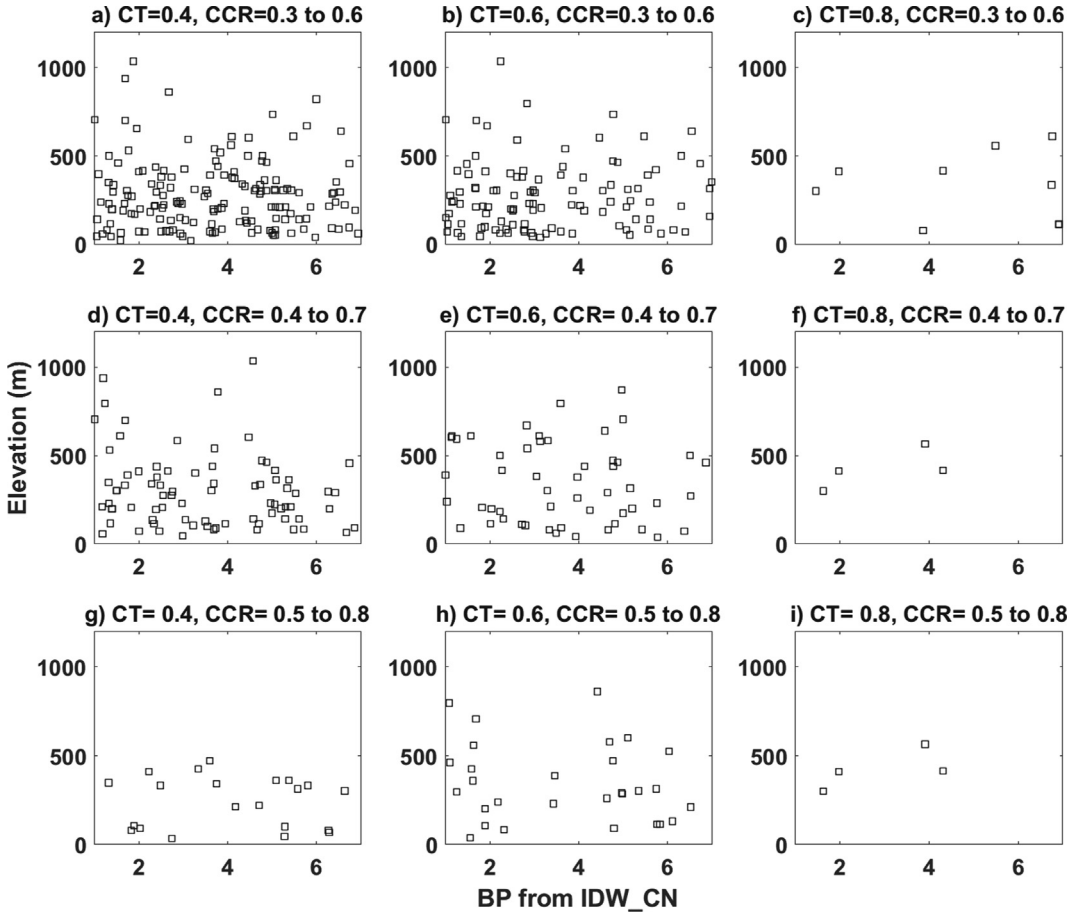

Fig. 11. Plots of elevation and BP associated with all valid stations for IDW_CN model.

Fig. 9(b) shows the performance of the IDW_CN model as a function of different elevation bands in which the rain gauges are located. For simplicity in representation, only CT = 0.6 and CCR = 0.3 to 0.6 are considered in this figure. The stations with lower RMSE and BP from the IDW_CN model than the IDW_NN model more or less equally belong to all elevation ranges. Therefore, it is not straightforward to derive any definitive conclusion about the effect of topography on the errors in reconstructing the rainfall using either the IDW_CN model or the IDW_NN model (Fig. 9(a) and (b)).

4.8. Variation of errors in IDW_CN with elevation of valid stations

Finally, we explore the relationship between the elevations of the valid stations with the associated error estimates from the IDW_CN models in reconstructing the rainfall data. In Figs. 10 and 11, RMSE and BP values associated with each of the valid station is plotted against its elevation for CT = 0.4, 0.6 and 0.8. Fig. 10 indicates that the RMSE is low at low elevations and increases as the elevation increases for all CT and CCR. The plot of BP associated with a valid station against the elevation of station do not show any particular trend (see Fig. 11), which indicates that the associated BP value is independent of the elevation of a valid station.

5. Discussion and conclusions

The results from our analysis show that there is no particular model, among IDW_NN, IDW_LN and IDW_CN, that has consistently the lowest RMSE and BP associated with all the valid stations. The IDW_LN model and the IDW_CN model show lower RMSE at about 30% of the stations; these two models have comparable (the difference is small) RMSE values at the rest 70% stations (see Tables 2, 3, and 4). Similarly, 50% of the stations show lower BP, while having no major changes in BP values at the remaining 50% of the stations. These results indicate that at least at around 30% of the stations in the IDW_LN model and the IDW_CN model show better results for rainfall reconstruction.

The effect of topography and elevation of valid stations on the performance of the three models can be interpreted from Figs. 5 and 9. As these figures show, the performance of the models is mostly independent of the location of the rain gauge stations. The effect of elevation of the rain gauge stations on the performance of the models can be explored further by considering a rain gauge network in which the rain gauges are distributed evenly in all elevation bands (unlike the MDB). The plots of the elevation of the rain gauge station versus the magnitude of the errors (RMSE and BP) (Figs. 10 and 11) show that the magnitude of RMSE associated with the IDW_LN model and the IDW_CN model has low values at low elevations and it increases as the elevation increases, whereas the associated BP is independent of the elevation of the rain gauge station. This indicates that the performance of the IDW_NN, IDW_LN and IDW_CN models depend upon the type of error statistics being considered.

For a natural system, traditional IDW approaches (such as the IDW_NN model) may be better suited than the networks-based approach (IDW_LN and IDW_CN), from the perspective of error statistics. However, they may not be completely effective in accounting the spatial rainfall variability, especially when the complexity of the system is high, such as where there are significant variations in topography and elevations. For such situations, the concepts of networks may be better suited, even when they are not able to completely explain the properties of the whole system. It is possible that a large part of a basin behaves in accordance with the spatial correlation assumption inherent in the

traditional approaches, while the other parts may have properties that may be better explained using the concepts of networks. As seen from the results of the present study, at some (valid) stations in the MDB, the traditional IDW_NN model provided very high errors, likely due to the significant variations in topography and elevations of the surrounding conditions. Therefore, using the new concept of networks (IDW_LN and IDW_CN models) along with traditional spatial correlation (IDW_NN) approach could very likely yield a much better methodology for the reconstruction of rainfall and other meteorological variables. The findings from this study are certainly helpful in filling the gaps in meteorological data, developing interpolated surface by carefully selecting the valid stations and classifying a region (catchment), among others.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research has been completed thanks to the support of the Science and Engineering Research Board (SERB), project number CRG/2018/000649 awarded to Sanjeev Kumar Jha. We thank two anonymous reviewers for their comments and suggestions on an earlier version of the manuscript, which has helped improve the quality and presentation of our work.

References

Barabási, A., Albert, R., 1999. Emergence of scaling in random networks. science.sciencemag.org.

Barros, A.P., Lettenmaier, D.P., 1993. Dynamic modeling of the spatial distribution of precipitation in remote mountainous areas. Mon. Weather Rev. 121, 1195–1214.

Barry, R.G., 1992. Mountain Weather and Climate [WWW Document]. Psychol. Press.

Berndtsson, R., 1988. Temporal variability in spatial correlation of daily rainfall. Water Resour. Res. 24, 1511–1517. https://doi.org/10.1029/WR024i009p01511.

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, 4386–4392. https://doi.org/10.1002/grl.50681.

Boers, N., Donner, R.V., Bookhagen, B., Kurths, J., 2015. Complex network analysis helps to identify impacts of the El Niño Southern Oscillation on moisture divergence in South America. Clim. Dyn. 45, 619–632. https://doi.org/10.1007/s00382-014-2265-7.

Buytaert, W., Celleri, R., Willems, P., Bièvre, B. De, Wyseure, G., 2006. Spatial and temporal rainfall variability in mountainous areas: A case study from the south Ecuadorian Andes. J. Hydrol. 329, 413–421. https://doi.org/10.1016/j.jhydrol.2006.02.031.

Daly, C., Halbleib, M., Smith, J.I., Gibson, W.P., Doggett, M.K., Taylor, G.H., Curtis, J., Pasteris, P.P., 2008. Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. Int. J. Climatol. 28, 2031–2064. https://doi.org/10.1002/joc.1688.

Di Piazza, A., Conti, F. Lo, Noto, L.V., Viola, F., La Loggia, G., 2011. Comparative analysis of different techniques for spatial interpolation of rainfall data to create a serially complete monthly time series of precipitation for Sicily, Italy. Int. J. Appl. Earth Obs. Geoinf. 13, 396–408. https://doi.org/10.1016/j.jag.2011.01.005.

Estrada, E., 2012. The Structure of Complex Networks: Theory and Applications [WWW Document]. Oxford Univ. Press.

Girvan, M., Newman, M.E.J., 2002. Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA 99, 7821–7826. https://doi.org/10.1073/pnas.122653799.

Griffith, D.A., 1992. What is spatial autocorrelation? Reflections on the past 25 years of spatial statistics. Espace. Geogr. https://doi.org/10.2307/44381737.

Jha, S.K., Sivakumar, B., 2017. Complex networks for rainfall modeling: spatial connections, temporal scale, and network size. J. Hydrol. 554, 482–489. https://doi.org/10.1016/j.jhydrol.2017.09.030.

Jha, S.K., Zhao, H., Woldemeskel, F.M., Sivakumar, B., 2015. Network theory and spatial rainfall connections: an interpretation. J. Hydrol. 527, 13–19. https://doi.org/10.1016/j.jhydrol.2015.04.035.

Konapala, G., Mishra, A., 2017. Review of complex networks application in hydroclimatic extremes with an implementation to characterize spatio-temporal drought propagation in continental USA. J. Hydrol. 555, 600–620. https://doi.org/10.1016/j.jhydrol.2017.10.033.

Li, T., Wang, G., Chen, J., 2010. A modified binary tree codification of drainage networks to support complex hydrological models. Comput. Geosci. 36, 1427–1435. https://doi.org/10.1016/j.cageo.2010.04.009.

Li, Z., Yang, D., Hong, Y., Zhang, J., Qi, Y., 2014. Characterizing spatiotemporal variations of hourly rainfall by gauge and radar in the mountainous three gorges region. J. Appl. Meteorol. Climatol. 53, 873–889. https://doi.org/10.1175/JAMC-D-13-0277.1.

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, 971–987. https://doi.org/10.1007/s00382-011-1156-4.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U., 2002. Network motifs: simple building blocks of complex networks. Science (80-.) 298, 824–827. https://doi.org/10.1126/science.298.5594.824.

Naufan, I., Sivakumar, B., Woldemeskel, F.M., Raghavan, S.V., Vu, M.T., Liong, S.Y., 2018. Spatial connections in regional climate model rainfall outputs at different temporal scales: application of network theory. J. Hydrol. 556, 1232–1243. https://doi.org/10.1016/j.jhydrol.2017.05.029.

Newman, M., 2012. Networks: An Introduction. 2010: Oxford Univ. Press. Artif. Life 241–242.

Robertson, J.C., 1967. The symap programme for computer mapping. Cartogr. J. 4, 108–113. https://doi.org/10.1179/caj.1967.4.2.108.

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

Shi, H., Chen, J., Li, T., Wang, G., 2017. A new method for estimation of spatially distributed rainfall through merging satellite observations, raingauge records, and terrain digital elevation model data. J. Hydro-Environment Res. https://doi.org/10.1016/j.jher.2017.10.006.

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, F.M., 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, F.M., 2014. Complex networks for streamflow dynamics. Hydrol. Earth Syst. Sci. Discuss.

Thiessen, A.H., 1911. Precipitation averages for large areas. Mon. Weather Rev. 39, 1082–1089.

Tobler, W.R., 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. Econ. Geogr. 46, 234. https://doi.org/10.2307/143141.

Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of 'small-world' networks. Nature 393, 440–442. https://doi.org/10.1038/30918.