Received: 23 August 2018 | Revised: 20 January 2019 | Accepted: 2 April 2019 | Published on: 24 April 2019

DOI: 10.1002/joc.6091

RESEARCH ARTICLE

International Journal of Climatology | RMetS

A conceptually superior variant of Shepard's method with modified neighbourhood selection for precipitation interpolation

Subash Yeggina[1] | **Ramesh S. V. Teegavarapu**[2] | **Sekhar Muddu**[1,3,4]

[1]Department of Civil Engineering, Indian Institute of Science, Bangalore, India

[2]Department of Civil, Environmental and Geomatics Engineering, Florida Atlantic University, Boca Raton, Florida

[3]Interdisciplinary Centre for Water Research, Indian Institute of Science, Bangalore, India

[4]Indo-French Cell for Water Sciences, Indian Institute of Science, Bangalore, India

Correspondence
Sekhar Muddu, Department of Civil Engineering, Indian Institute of Science, Bangalore, India.
Email: sekhar.muddu@gmail.com

Abstract

The accuracy of gridded precipitation data depends on the availability of a uniformly spaced rain gauge network and an appropriate spatial interpolation method that considers the rainfall variability and other factors that influence the precipitation patterns in the region of interest. In the current study, conceptually superior variants of a widely used spatial interpolation algorithm, Shepard's method, are proposed, formulated and evaluated to overcome one of the major limitations in neighbourhood selection, that is, arbitrary selection of rain gauges. The variants provide mechanisms to objectively select the rain gauges (control points) based on correlation (variant 1), distribution similarity (variant 2) and a combination of both (variant 3). The improved variants were used in the development of gridded rainfall data at a resolution of 5 km over the Kabini River basin in south India, and in the state of Kentucky, United States. Results from multiple experiments using the original Shepard's method and its variants indicate improvements in the accuracy of precipitation estimates. Also, these variants have preserved the site-specific statistics and distributional characteristics of the rainfall data. A variant 1 that uses a correlation-based neighbourhood selection criterion performed better for daily and monthly data compared to others and is suitable for generation of gridded rainfall data. The variant 1 when used with information from clustering of sites for selection of the neighbours has led to improvement in gridded precipitation data estimates. The proposed variant 1 can also be used for point data estimation useful for filling missing data at any site.

KEYWORDS

clusters, distribution similarity, gridded rainfall, missing data, neighbourhood selection, Shepard's method, spatial interpolation

1 | INTRODUCTION

Accurate gridded precipitation data are essential for hydrological modelling, climate change and attribution studies, evaluation of climate model simulations, and assessment and bias correction of satellite rainfall products. Generation of fine-scale gridded rainfall data from point observations is quite challenging in the Tropics, due to extreme variability in intensities, seasonality, complexity in the terrain (Malhi and Wright, 2004) besides sparse and non-homogeneously distributed gauge network (Hofstra *et al*., 2009).

Several spatial interpolation techniques such as inverse distance weighting (IDW) method, thin plate splines (TPS), trend surface analysis (including linear and polynomial regression), kriging and parameter elevation regression on independent slopes model (PRISM; Hutchinson, 1995;

American Society of Civil Engineers (ASCE), 1996; Goovaerts, 1997; Hay *et al*., 1998; Daly *et al*., 2002) were used either independently or in a combination (Lin and Chen, 2004; Perry and Hollis, 2005; Camera *et al*., 2014) to generate gridded precipitation data in the last two decades. Among those, the deterministic IDW and the kriging methods with their respective variants are widely used both for the estimation of missing rainfall records and interpolation (Teegavarapu, 2007; Li and Heap, 2011; Ly *et al*., 2011; Yanto *et al*., 2017). Although kriging and its variants such as universal kriging (UK) and co-kriging (CK) which can incorporate covariates in the prediction are well documented and used in the generation of gridded rainfall, the estimation accuracy of these methods depends on the fitting of a best empirical variogram for the spatial structure of the observed variable (Burrough and McDonnell, 1998). Also, the identified spatial structure based on empirical variogram may not always conform to the overall spatial structure of the observed variable due to several reasons (e.g., poor quality of data or presence of spatial heterogeneity in the data). Furthermore, the difficulty in selection of the best variogram model and assumptions of normality and stationarity of the data set limits its utility. Also, a spatially uniform observational data set (e.g., rain gauge network) is also needed. Copula-based methods have been used successfully for spatial interpolation in the past decade. Bárdossy and Li (2008) and Bárdossy (2011) have applied a copula-based approach for spatial interpolation as an alternate to spatial modelling using the variogram and demonstrated its applicability for ground water quality parameters. Wasko *et al*. (2013) also used a copula-based approach for combining local and global spatial predictions. The selection of nearby stations used to calculate the copula density is based on distance.

IDW is an empirical deterministic exact interpolation method, which is easy to implement and sometimes does outperform sophisticated approaches such as kriging (Ruelland *et al*., 2008). It can perform well compared to regression-based interpolation in the generation of gridded daily rainfall data and when used as forcing in a distributed hydrological model as reported in a study by Hwang *et al*. (2012). However, one of the major limitations of IDW is that it does not consider the spatial distribution of the observations available at different control points. Shepard (1968; 1984) modified the IDW using angular and distance-based weights to account for the directional distribution of data in space. A widely used interpolation tool, Synergraphic Mapping System (SYMAP), adopted the approach presented by Shepard. The Shepard's method (SM) is also referred to as angular distance weighting (ADW) interpolation method (New *et al*., 2000). Chen *et al*. (2002) reported a comparative study of three interpolation techniques for precipitation, namely Cressman, Shepard and a statistical method, optimal interpolation (OI) of Gandin over global land areas and found that the performance of SM is comparable with that of OI technique for different rain gauge network densities. Several variants to the weights of the original SM were predominantly developed in various past studies. Willmott *et al*. (1985) adapted the method originally designed for Cartesian space to spherical surface and referred to it as the SPHEREMAP (Spherical Mapping System). Frei and Schär (1998) modified the distance-based weights by adopting a smoother radial weighting function. Simolo *et al*. (2010) used the elevation differences as additional weights in missing data imputation. New *et al*. (2000) modified the distance-based weighting function in SM with an empirical function based on correlation and distance referred to as correlation distance decay (CDD). The CDD helped in defining the search radius which was arbitrary in the original SM, and also to derive a weight for each station included in the grid value estimate (Hofstra and New, 2009).

Numerous studies in the past have created global and regional gridded climate data using the variants of SM. Among them, the most popular is the global land surface precipitation data set Global Precipitation Climatology Centre (GPCC) developed using the SPHEREMAP interpolation method at a monthly time interval and 0.5° spatial resolution (Rudolf and Schneider, 2005). Similarly, several gridded products of climatic variables such as temperature, precipitation, cloud cover, vapour pressure, and others were developed using ADW interpolation method at 0.5° spatial resolution and monthly temporal scale for global land mass (Mitchell and Jones, 2005; Harris *et al*., 2014). These products are referred to as Climatic Research Unit (CRU) Time-Series (TS) data version 4.01. Likewise, ADW is used to develop the University of Delaware (UDel) precipitation and temperature global gridded, high-resolution monthly data sets. The APHRODITE (Asian Precipitation Highly Resolved Observational Data Integration Towards the Evaluation of Water Resources) is also created using the modified SPHEREMAP to interpolate precipitation data at various resolutions (Yatagai *et al*., 2012). Livneh *et al*. (2013) generated a long-term (1915–2011), 97-year precipitation and temperature gridded daily data set at 0.0625° resolution (approximately 6 km) for conterminous United States, using the SYMAP and corrected it by applying the PRISM algorithm to match the long-term mean of the gridded precipitation data, to benefit energy exchanges, water balances, and climate change impact studies.

Meterological products available from national agencies at country level have used SM for griddded data generation. The Indian Meteorological Department (IMD) developed multiple gridded daily rainfall products (Rajeevan *et al*., 2005; Pai *et al*., 2014) at different spatial resolutions (1, 0.5, and 0.25°) using SM over the Indian land area. Srivastava

et al. (2009) developed a daily gridded temperature data at 1° resolution over whole India using ADW, by selecting the neighbourhood based on CDD. Xavier *et al*. (2015) compared six interpolation methods for gridded data generation of six weather variables at daily and monthly scales at 0.25° spatial resolution over Brazil and found that IDW and ADW were the best methods for all the variables. Nguyen *et al*. (2016) generated the Vietnam-Gridded Precipitation (VnGP) data using Shepard's interpolation method at a daily scale for two spatial resolutions of 0.25 and 0.1° from 1980 to 2010.

The definition of the neighbourhood and ultimately the selection of the stations which will be used in the interpolation are crucial to the success of any spatial interpolation method in providing accurate estimates of the variable of interest. The way the inter rain gauge relationship is characterized is different for each interpolation approach. The standard IDW method expresses this relationship using a geographical distance (i.e., Euclidean distance) whereas ordinary kriging and its variants use variogram models that are typical implementations of monotonic functions of geographical distance (Ahrens, 2006). Several variants to geographical distance for selection of neighbouring stations were conceived and evaluated in the earlier studies particularly for missing data generation. Ahrens (2006) used a statistical measure based on a mean squared difference of precipitation timeseries between the reference station and the neighbouring stations and found it to be suitable for application in mountaineous and regions with high rainfall variability. Similarly, a statistical parameter considering the rainfall data measurement uncertainty was developed for selecting the neighbourhood stations in filling monthly missing data (Ramos-Calzado *et al*., 2008).

Simolo *et al*. (2010) proposed a two-step imputation method, where a wet or dry day probability at the reference station is first estimated from the neighbouring stations and the precipitation amount is estimated by multiple linear regression, to ensure that probability distribution and long-term statistics are preserved. Teegavarapu and Chandramouli (2005) argued that distance alone is not the best measure of spatial autocorrelation. They proposed several conceptual improvements to IDW method for estimating missing data. One such improvement was the selection of nearby stations based on the strength of correlation quantified by Pearson coefficient of correlation (CC) calculated based on observations from the reference site and any other site. They found these improvements to IDW method to be superior to distance-based weighting methods. Considering the improvements observed in missing data filling by selection of neighbours based on different criteria, in the current study different variants of SM are proposed and evaluated for selection of the neighbourhood for generation of gridded data.

Although SM is a robust spatial interpolation method, guidelines or rules for neighbourhood selection (i.e., selection of a set of control points in space) are arbitrary or mostly framed based on experience. There are several shortcomings of the method, and they include (a) the neighbourhood is established arbitrarily by a distance-based criterion; (b) lack of guidance in the selection of neighbourhood points for interpolation in space and time; and (c) the method does not consider factors such as changes in the precipitation regimes due to land use and topographical and meteorological features in interpolation.

Considering the main limitation of SM (i.e., arbitrary neighbourhood selection), this study has proposed, developed, and evaluated conceptually improved procedures for objective selection of control points within the neighbourhood of the point of interest (i.e., grid centre) by correlation (variant 1), by distribution similarity (variant 2), and a combination of both (variant 3). The contents of the paper are organized as follows. First, SM is described, and formulations of the variants of the method related to the neighbourhood selection are presented. Case study details and data description are provided next, followed by applications of these variants, results and analysis, general remarks, and conclusions.

2 | INTERPOLATION METHOD

This section describes Shepard's interpolation method and provides the details of the selection of the neighbourhood and several conceptual improvements to it.

2.1 | Shepard's method

SM is an inverse distance weighted interpolation algorithm for irregularly spaced data points or control points (D_k). A continuous function where the weighted average of data is inversely proportional to the distance from the interpolated location or the point of interest is used in SM (Shepard, 1968). The classical form of IDW interpolation function is given by Equation (1),

$$\hat{P}_j = \sum_{k=1}^{n} w_k P_k \qquad \forall j, \tag{1}$$

where (x_k, y_k) is the location coordinates of D_k, whose values are represented by P_k. The control points are the observation locations selected for interpolation. Let (x_j, y_j) represents the coordinates of grid centre (GC) of the lattice where the values are to be estimated with $\hat{P}_j$ the estimated value at j. The weights are given by $w_k = \frac{d_{j,k}^{-\alpha}}{\sum_{k=1}^{n} d_{j,k}^{-\alpha}}$, where $d_{j,\,k}$ is the

Euclidean distance between the prediction location j and data point k and n is the number of nearby stations used in the estimation of value at GC_j. The variable α is the power parameter and a value of two is generally used in interpolation.

IDW tends to generate bull's eye patterns of concentric contours around the data points (Burrough and McDonnell, 1998). All interpolated values between data points lie within the range of the data point values and hence these values may not approximate valleys and peaks well. Shepard (1968) developed a local interpolant that improved the IDW through a three-step weighting mechanism to account for the proximity of control points not only by the Euclidean distance but also for angular distribution of data points considered in the estimation of the grid estimate as well as their spatial gradients within the data.

The interpolation function (Shepard, 1968) is given by

$$\hat{P}_j = \begin{cases} \dfrac{\sum_{k=1}^{n} W_k (P_k + \Delta P_k)}{\sum_{k=1}^{n} W_k}, & d_{j,l} > \epsilon \\ \dfrac{1}{m} \sum_{k=1}^{m} P_k, & d_{j,l} \le \epsilon \end{cases}, \quad (2)$$

where ϵ is a radius given by $\epsilon = 0.01 \max(\Delta x, \Delta y)$, Δx and Δy are the width and height of the grid cells. In this case, the grid being 5 km, $\epsilon = 50$ m. The variable m is the number of data points within a distance ϵ to j. If the data points are within the radius of ϵ then the average of their values is computed else the first part of the Equation (2) is used to estimate at grid j. ΔP_k is the adjusted data point value for slope correction.

The weight (W_k) for each control point given by

$$W_k = S_k^{\gamma}(1 + T_k) \quad \forall k, \quad (3)$$

where S_k is the distance-based weight and T_k is an angular weight for data point k. A value of two is used for the exponent γ (Shepard, 1968; 1984).

The weight based on distance (S_k) is defined using three non-overlapping distance related intervals and is given by Equation (4),

$$S_k = \begin{cases} d_{j,k}^{-1}, & d_{j,k} \le \dfrac{r_j}{3} \\ \dfrac{27}{4 r_j} \left(\dfrac{d_{j,k}}{r_j} - 1 \right)^2, & \dfrac{r_j}{3} < d_{j,k} \le r_j \\ 0, & d_{j,k} > r_j \end{cases} \quad \forall j,k, \quad (4)$$

where r_j is the search radius.

A directional weighting (T_k) for each data point k is given by Equation (5),

$$T_k = \frac{\sum_{l=1}^{n} S_l \left[1 - \cos \theta_j(k,l) \right]}{\sum_{l=1}^{n} S_l} \quad \forall l \ne k, \quad \forall j,k, \quad (5)$$

where $\theta_j(k, l)$ is the angular separation between the two vectors with data point k, l and j as the vertex and its distance-based weight is S_l. The angle between two vectors $\cos\theta_j(k, l)$ is given by

$$\cos \theta_j(k,l) = \frac{(x_k - x_j)(x_l - x_j) + (y_k - y_j)(y_l - y_j)}{d_{j,k} d_{j,l}}. \quad (6)$$

The value of $\cos\theta_j(k, l)$ ranges between -1 and 1. The data which are clustered will have a small angular separation and contribute less, thus reducing their influence in the estimation, compared to the points having a large angular separation. The SM creates "flat-spots" at the data point locations, that is, the gradient is zero at all the data points. To overcome this undesirable behaviour, increments (ΔP_k) are computed at each D_k which is an average weighted gradient for each value of the D_k used in the interpolation, based on the collective rates of change at the other data points within r_j. This is added to the respective D_k values so that the surface would achieve desired partial derivatives. The readers may refer to Shepard (1968) and Willmott *et al.* (1985) for further details.

2.2 | Neighbourhood size selection

The neighbourhood size determines how many control points with observations are included in the interpolation for the grid estimate. A support set is defined as the set of data points included in the estimation of a variable value at grid centre j. Figure 1a illustrates an example of how the support set is built. Shepard (1968) proposed two criteria for building the support set and they are:

1. An arbitrary distance criterion which selects all the points falling within r_j with j as the centre, such that

$$C_j = \{ D_k \mid d_{j,k} \le r_j \}; \text{cardinal number } n(C_j) = N \quad \forall j, \quad (7)$$

where N is the total number of data points.
Next D_k in set C_j is ordered in increasing distance from j

$$d_{j,k} = d_{j,1} \le d_{j,2} \le \ldots \le d_{j,N} \quad \forall j. \quad (8)$$

2. An arbitrary number of data points are chosen which further reduces the set to only the select number. If n is the

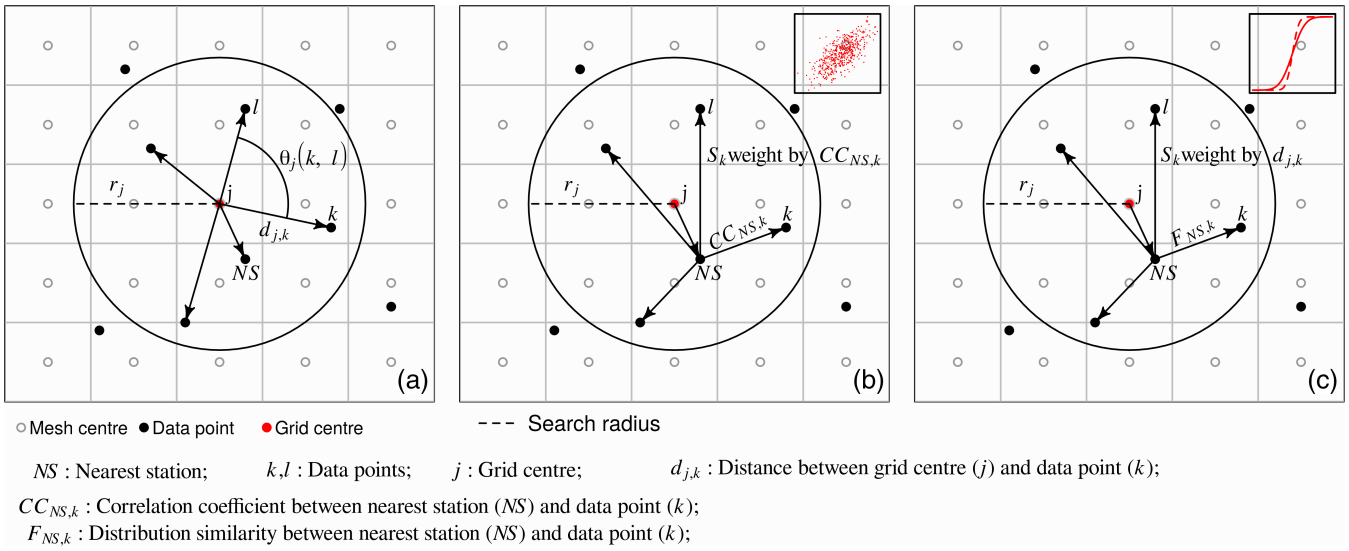

FIGURE 1 Schematic illustration of neighbourhood selection using (a) distance, (b) variant 1 (using correlation; inset showing the correlation between data point NS and *k*), (c) variant 2 (using distribution; inset showing the CDF of data point NS [solid line] and *k* [dotted line]) [Colour figure can be viewed at wileyonlinelibrary.com]

chosen number of D_k then the support set is given by Equation (9),

$$C_j^n = \{D_1, D_2, D_3, \ldots D_n\} \mid n \leq N \qquad \forall j, n; n \leq N. \tag{9}$$

The second criterion presumes that a defined subset of interpolating points was optimal, regardless of the relative location and spacing of the points. Shepard suggested that applying the two criteria together has their advantages limiting to a minimum of 4 data points and a maximum of 10 data points to be included in the interpolation. Accordingly, r_j needs to be adjusted to meet the two criteria. The neighbourhood thus selected is referred to as $\mathrm{SM}_{\mathrm{NN}_d}$, where NN are the nearest neighbours and *d* is the distance. The variants proposed and evaluated in this study are discussed in the following sections.

2.3 | Variant 1: Use of correlation

Unlike filling the missing records at a rain gauge, where data are available at the reference station, in the case of gridded data generation, there is no reference time series to develop the relationship. Therefore, in the current study, it is assumed that the nearest station (NS) to the centre of the grid at which the value of the estimate is sought could be the best suitable candidate which has similar characteristics of where the data are being interpolated. All the gauges within a prespecified boundary are evaluated based on the correlation with the NS and ranked based on the correlation coefficient value from highest to lowest, and a specific number of highest ranked (by correlation) sites are selected for interpolation. This variant is referred to a $\mathrm{SM}_{\mathrm{NN}_\rho}$, where ρ is the correlation. An illustration explaining the variant 1 is shown in Figure 1b. The correlation between the observations at NS and each site within C_j is referred to as $\mathrm{CC}_{\mathrm{NS},k}$ and $k \leq N-1$ given by Equation (10),

$$\mathrm{CC}_{\mathrm{NS},k} = \mathrm{CC}_{\mathrm{NS},2}, \mathrm{CC}_{\mathrm{NS},3}, \ldots, \mathrm{CC}_{\mathrm{NS},\mathrm{N}-1} \qquad \forall k, k \leq N-1. \tag{10}$$

Next, D_k in Equation (8) are sorted in increasing order of correlation and *n* points in D_k are chosen as the support set $\left(C_j^{n'}\right)$ given by Equation (11),

$$C_j^{n'} = \{D_{\mathrm{NS}}, D_2, D_3, \ldots, D_n\} \mid n \leq N \qquad \forall j, n. \tag{11}$$

A modification to the distance-weighting scheme in Equation (4) is required, although the use of variant 1 selected suitable stations for interpolation as SM weights the stations only based on distance, some of the stations which are not in the SM support set but in the $\mathrm{SM}_{\mathrm{NN}_\rho}$ support set will get lower weights, using weights based on distance (S_k) as in Equation (4). Therefore, a correlation-based weight is introduced in this variant. The corresponding correlations derived between the NS and all the stations are used as weights in the estimation, with NS assumed a weight of 1 (for correlation − highest correlated site). The modification to Equation (4) is provided as Equation (12), where *k* is set of D_k in the support set,

$$S_k = \frac{1}{e^{(1-CC_k)}} \qquad \forall k. \tag{12}$$

Although it is conceptually possible to have different weighting functions similar to distance weights based on search radius in Equation (4) or different classes based on the correlation strength, subjectivity in choosing the thresholds for the classes is the main limitation. Therefore, in the current study, we used a single weighting function. The other equations for angular weights and gradient correction are the same as described in section 2.1.

2.4 | Variant 2: Use of distribution similarity

A new criterion for neighbourhood selection is proposed in this study such that all the gauges within the pre-specified boundary are evaluated based on distribution similarity to the NS. The illustration is shown in Figure 1c. The two-sample Kolmogorov–Smirnov (KS) test is used to decide whether two random samples have the same statistical distribution. The KS statistic quantifies the distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples. The null hypothesis of the KS test is that both distributions are identical, without any further assumption regarding their location and shape, which makes the KS test widely applicable. This variant is used as additional criterion for selecting the neighbourhood where the set of D_k in Equation (8) are tested for their distribution similarity as given in Equation (13),

$$F_{\mathrm{NS},k} = F_{\mathrm{NS},2}, F_{\mathrm{NS},3}, \ldots, F_{\mathrm{NS},\mathrm{N}-1} \quad \forall k,\ k \leq N-1. \quad (13)$$

Next D_k in set C_j^n are the data points which have passed the KS test and given by

$$C_j^{n''} = \left\{ D_{\mathrm{NS},2}, D_{\mathrm{NS},3}, \ldots, D_{\mathrm{NS},n-1} \right\} \mid n \leq N \quad \forall j,n, \quad (14)$$

where $C_j^{n''}$ is the support set containing the distribution similarity of data points with the NS.

This variant is referred to as $\mathrm{SM}_{\mathrm{NN}_F}$, where the use of subscript F refers to the variant with distribution similarity.

2.5 | Variant 3: Use of correlation and distributional similarity

The variant 3 is a combination of variants 1 and 2. The gauges are ordered based on the strength of correlation to NS, and further only those which are distributionally similar to the NS are selected in the interpolation process. This variant is referred to as $\mathrm{SM}_{\mathrm{NN}\rho F}$.

3 | METHODOLOGY

The interpolation methodology is described in a step-by-step approach that is generic to gridded data generation and also applicable for filling missing data and is shown in Figure 2. The steps are explained as follows:

1. Once an interpolation location (j) is defined, a typical interpolation goes with defining r_j. The set C_j is built.
2. The D_k are ordered based on the Euclidean distance ($d_{j,\,k}$). The closest station to j is designated as the NS.
3. The correlation either parametric or nonparametric is established between the NS and other data points D_k in the set C_j. The data points are re-ordered with respect to correlation in the decreasing order.
4. Similarly, the distribution similarity is ascertained between the NS and other stations in r_j.
5. Therefore, selection of the stations to be used in interpolation can be picked from the set in 3 or from 4, or a combination of 3 and 4.
6. A definite number of data points used in interpolation is assumed to be n. The set is trimmed to n and this is the support set.
7. As the neighbourhood is defined, the gridded data can be generated or the missing data can be estimated.
8. The estimated values are compared and the best variant is chosen.

3.1 | Experiments with variants

The application and evaluation of the variants of neighbourhood selection developed in this study are carried out in a series of experiments. In the first experiment, the performance of SM is compared with that of variant 1 for Kabini River basin and the Kentucky regions. The second experiment is used to test the sensitivity of the data length on the performance of the variant 1. The third experiment tests the benefits of using a site NS that belongs to a set of sites defined by a clustering approach using the variant 1. The fourth experiment evaluates variant 1 using anomalies rather than absolute values of rainfall. The fifth experiment is used to evaluate the use of variant 2. The sixth experiment assesses the use of variant 3. The seventh experiment investigates the potential of using variant 1 for filling in missing data.

3.1.1 | Experiment 1: Effect of the size of the neighbourhood

An experiment is conducted to test how the error measures improve or worsen with the inclusion of more stations in the

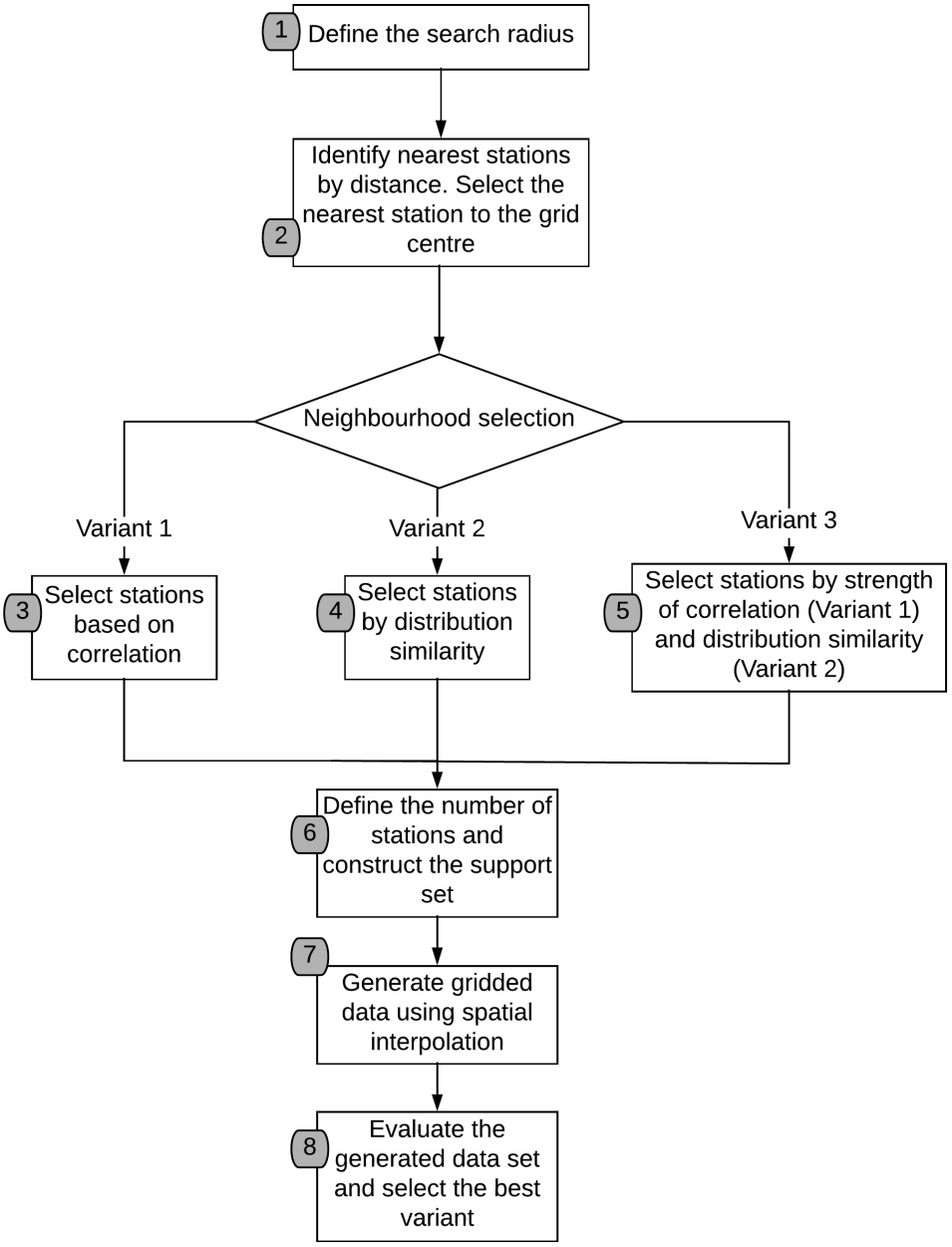

FIGURE 2 Workflow of gridded data generation using three variants of SM. Numerals represent the step described in the methodology

interpolation. To assess the performances of the estimation methods SM_{NN_d} and SM_{NN_ρ} with an increase in the number of stations used in interpolation, cross validation is carried out by gradually increasing the number of stations. The weight S_k for SM_{NN_d} used is defined in Equation (4) and for SM_{NN_ρ}, a weight based on correlation is provided by Equation (12).

3.1.2 | Experiment 2: Sensitivity of time series length in neighbourhood selection

An experiment is carried out by varying the length of the time series (TL) to determine if the support set changes with the increase or decrease in data length and evaluate how the error measure varied with the inclusion of more stations. The influence of TL on performance metric is evaluated using 1, 2, 5, 10, 15, and 30 years of data. Random samples are generated as a sequence of years, and each experiment was repeated 10 times.

3.1.3 | Experiment 3: Effect of choosing the NS in a climatic zone

The selection of the NS in Experiments 1 and 2 is critical, as the selection of NS determines which sites have the highest correlation scores and thus which sites are included in the neighbourhood. If the region has uniform rainfall variability, the selection of NS may be inconsequential. However, if there are heterogeneous rainfall areas in the region, then NS should be carefully selected. Selection of NS can be based

on homogeneous rainfall areas identified using a *k*-means clustering technique or delineated boundaries of homogeneous areas by visual assessment of summary statistics of rainfall data or topographical features or land use information. The selection is referred to as $SM_{NN_{barrier}}$. A *k*-means clustering approach (Wichern and Johnson, 1992) with Euclidean distance as a measure of similarity was used for rainfall regionalization and delineation of homogeneous zones in this study.

3.1.4 | Experiment 4: Interpolation using anomalies

Willmott and Robeson (1995) have suggested that instead of using the observed totals, an indirect approach of normalized values can be used to alleviate the climatological variations and leave a smoothly varying surface. This procedure is known as climatology aided interpolation (CAI), in which the normalization is carried out either by subtracting the long-term average from the observed totals, or by dividing the observed totals with the long-term monthly mean value and the interpolation is done in a two-step approach. In the first step, the long-term means of station rainfall were interpolated to the desired resolution, and this is used as a "First Guess" or monthly climatological field or background fields. In the second step, for example, for daily gridded data, the observed daily rainfall values are divided by the long-term average and these ratios are interpolated and later multiplied with the First Guess field (Hunter and Meentemeyer, 2005; Hiebl and Frei, 2017). In case of monthly data the anomalies (i.e., observed data minus long-term mean) at each station are interpolated (i.e., monthly anomaly field) using TPS and are added back to the pertinent monthly climatological field (Chen *et al*., 2002; Perry and Hollis, 2005; Isotta *et al*., 2014).

3.1.5 | Experiment 5: Neighbourhood selection using distribution similarity

The distribution similarity between the NS and the surrounding stations in the neighbourhood is proposed as a new criterion for selection of gauges after the neighbourhood stations are selected by distance.

3.1.6 | Experiment 6: Neighbourhood selection using correlation and distribution similarity

The stations are first selected based on the correlation, and further, they are checked for their distribution similarity with the NS. This variant is tested with the monthly rainfall data for the Kabini River basin.

3.1.7 | Experiment 7: Filling missing data

This experiment was done to test the performance of variant 1 in estimation of daily missing precipitation data as several studies (e.g., Teegavarapu and Chandramouli, 2005; Teegavarapu, 2014) in the past applied both deterministic and stochastic spatial interpolation methods for reconstruction of missing data at one single site. There are two scenarios in missing data imputation. In the first scenario the relationships are established with reference to base station (BS; i.e., station or site at which missing data are known to exist) and the nearby stations within the search radius and the control points are identified. In the second scenario the NS to the BS is identified and the corresponding control points are ascertained, and missing data is gap filled. The experiment was conducted assuming that approximately 33% of historical data (15 years) is missing.

3.2 | Error and performance measures

The proposed variants of SM in this study are evaluated using a number of error and performance measures. The performance evaluation of each variant in the neighbourhood selection is carried out using a leave-one-out cross validation (LOOCV) approach. LOOCV is a common approach used to evaluate the interpolation results, where the data for the test site is removed from the original data set and is estimated using the interpolation methods. Common error measures such as mean error (ME) also referred to as bias, root-mean-square error (RMSE), performance metric of CC and index of agreement (*d*) (Li and Heap, 2011) are used to compare the observed data (P_i) with the estimated value ($\hat{P}_i$) at the test site (*i*). The index of agreement (*d*) proposed by Willmott (1981) is given by

$$d = 1 - \frac{\sum_{i=1}^{L} \left(P_i - \hat{P}_i\right)^2}{\sum_{i=1}^{L} \left(\left|\hat{P}_i - \overline{\hat{P}}_i\right| + \left|P_i - \overline{P}_i\right|\right)^2}, \tag{15}$$

where $\overline{P}_i$ is the mean of the observed data, $\overline{\hat{P}}_i$ is the mean of the estimated values, for a given interval *i* and *L* is the length of timeseries. The *d* values range between 0 and 1 with higher index values indicating that the estimated values agree well with the observations.

3.3 | Check for distributional similarity

Although the error measures and performance metrics would help in deciding the best interpolation method, for a given region, exhaustive validation to check if distribution, autocorrelation, and spatial structure are preserved in the interpolated data can be beneficial. In the initial assessment stage,

the observed and estimated daily and monthly rainfall data distributions are compared visually using Kernel density estimate (KDE) plots. Skewness and kurtosis of the observed and estimated data sets are also compared in this study.

3.3.1 | KS test

The two-sample KS test, a nonparametric test, is used to test the agreement between the cumulative distributions of observed and estimated precipitation data. The test is known to be sensitive to differences between data sets that may not be apparent through comparison of summary statistics (Wilcox, 2005) or by visual comparison of cumulative distribution function plots. The KS test checks for any violation of the null hypothesis for different medians, different variances, or different distributions. The null hypothesis that the two samples were drawn from the same distribution is rejected at the chosen significance level if a test statistic, is greater than a pre-specified critical value. In the current study a 5% significance level is chosen. This test was conducted to check if the observed and the predicted data at a station come from the same population considering only non-zero precipitation. A variant of the KS test (Sekhon, 2011) that considers ties in the sample data sets is used in this study.

3.4 | Variogram analysis: Evaluation of spatial dependence

The distance–decay relationship within the spatial data is characterized typically by a variogram (Cressie, 1993). A variogram can help to evaluate whether the spatial structure as observed from the gauge-based rainfall is also preserved in the generated gridded rainfall data. A variogram is a measure of how observations vary with increasing separation distance and is calculated by

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{N(h)} [P_i - P_k]^2, \tag{16}$$

where $\hat{\gamma}(h)$ is the semi-variogram value. P_i and P_k are the precipitation at spatial locations (x_i, y_i) and (x_k, y_k), respectively. $N(h)$ is the number pairs of data points separated by a Euclidean distance, $i - k = h$.

3.5 | Persistence evaluation

The persistence in time series of precipitation data is evaluated using serial autocorrelation. To evaluate how well the gridded precipitation data sets preserve the serial autocorrelation at different temporal lags, the serial correlation based on the observed data and the estimated data are determined and compared.

4 | STUDY DOMAIN AND DATA

4.1 | Study area

The variants of the SM are evaluated for gridded data generation using rain gauge data from two diverse climatic regions in different parts of the world. These regions are (a) Kabini River basin in the southwestern part of India and (b) the state of Kentucky, United States.

4.1.1 | Kabini River basin, India

The Kabini River basin in the peninsular part of India is on the leeward side of the Western Ghats and lies between 11°30′9″N to 12°21′22.68″N latitude and 75°47′25.44″E to 76°54′37.44″E longitude with a distinct climate and morpho-pedologic gradient (Figure 3a). The west–east geomorphologic gradient is due to the climatic gradient induced by the Western Ghats, which run parallel to the west coast and act as a barrier to the monsoon winds. Due to its unique characteristics, the basin has been identified as a Critical Zone Observatory (CZO) where multidisciplinary studies related to hydrology (Subash *et al*., 2017), geochemistry (Buvaneshwari *et al*., 2017), agro-hydrology (Sreelash *et al*., 2017), soil moisture (Tomer *et al*., 2015), remote sensing of vegetation (Eswar *et al*., 2017a; 2017b; Sharma *et al*., 2018), and eco-hydrology (Chitra-Tarak *et al*., 2018) have been conducted (Sekhar *et al*., 2016) in the recent past. The area of the watershed is approximately 7,000 km^2. The elevation ranges from 500 to 2,000 m above MSL. Figure 4a–d shows the spatial fields of four rainfall moments generated by cubic spline interpolation using *interp* function in the R package *fields* (Nychka *et al*., 2017). There is a high gradient in rainfall with the mean annual rainfall varying from 5,000 mm in the west to 800 mm to the east (Figure 4a). The temporal variability in rainfall is also quite high. The rainfall data distribution is right skewed at most of the stations in the region. The basin falls into two different climatic zones: tropical monsoon and tropical savannah according to Köppen–Geiger classification (Kottek *et al*., 2006).

4.1.2 | Kentucky, United States

The state of Kentucky extends between 36°29′49.56″N to 39°8′51.72″N latitude and 81°57′54″W to 89°34′16.32″W longitude. The state of Kentucky covers 1,04,623 km^2 of territory (Figure 3b). Most elevations are less than 107 m above sea level, with its lowest elevation at approximately 76 m above sea level (Kleber, 2015). The climate is humid

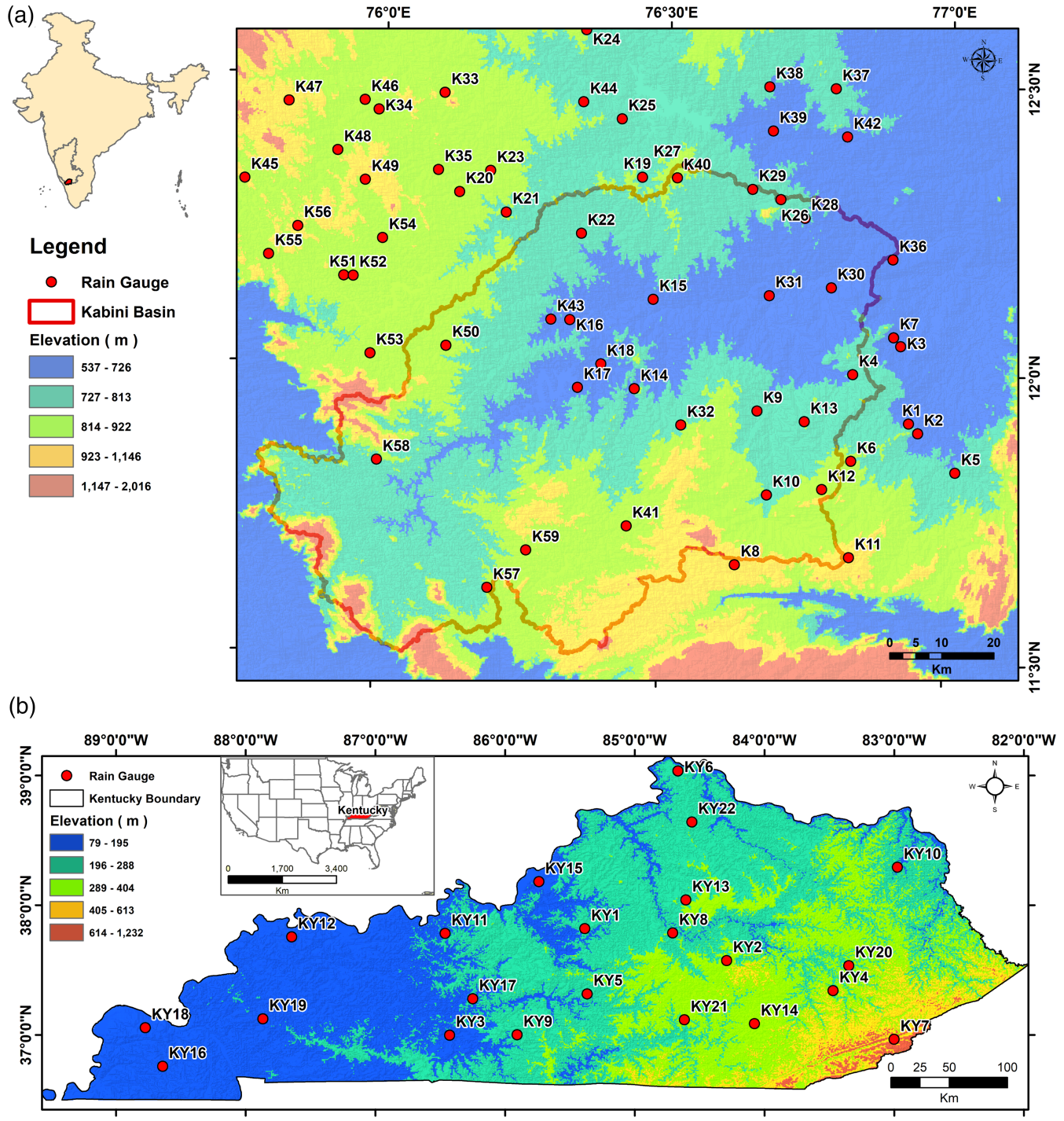

FIGURE 3 Elevation map of study area with the rain gauge network: (a) Kabini River basin and (b) Kentucky state [Colour figure can be viewed at wileyonlinelibrary.com]

subtropical with the annual average precipitation over the state varying from 1,060 mm in the north to 1,502 mm in the southwest with an average annual temperature ranging from 10.8°C in the northeast to 14.1°C in the southwest (Kleber, 2015). The state of Kentucky is classified as temperate, fully humid and hot summer type of climate according to Köppen–Geiger classification scheme (Teegavarapu, 2014). Figure 4e–h shows the spatial fields of four statistical moments of rainfall data. The temporal variability in rainfall is high in the southwestern region of the state. The rainfall data distribution in the central part of the region is symmetric and skewed (i.e., switching from left to right skewness) in the southern region as shown in Figure 4g.

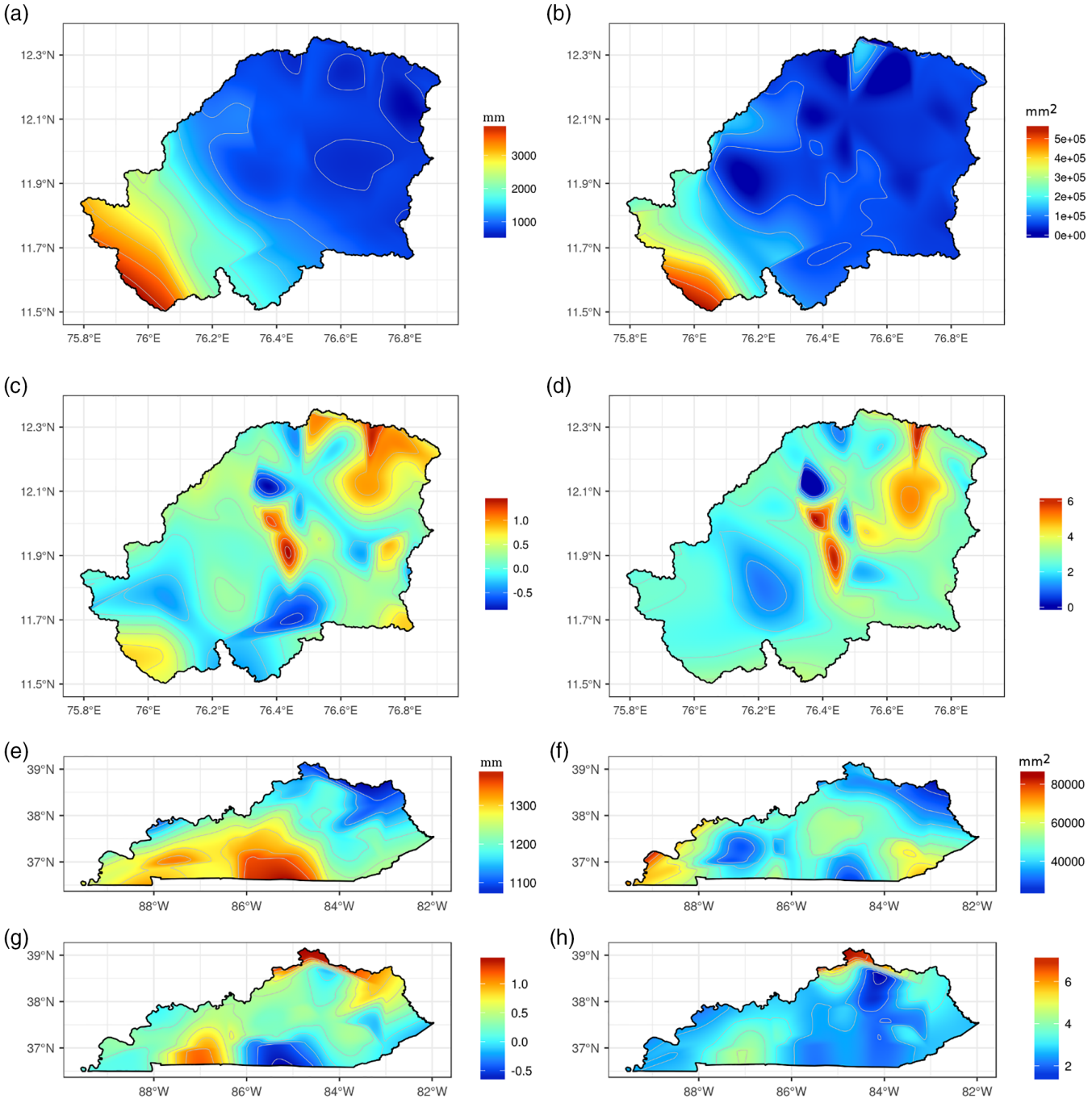

FIGURE 4 Spatial variation of four moments of rainfall distribution for Kabini River basin (top four panels) and Kentucky state (bottom four panels). Mean (a, e), variance (b, f), skewness (c, g) and kurtosis (d, h) [Colour figure can be viewed at wileyonlinelibrary.com]

4.2 | Description of data sets

Daily rain gauge data for Kabini River basin used in this study were collected from the Department of Economics and Statistics (DES), Karnataka, and the IMD, Thiruvananthapuram. The locations of the gauges used in this study are shown in Figure 3a and are numbered as K# for convenience. There are 59 gauges in the study region with a record length of 30 years (1980–2009), and the percentage of missing data length varied from 3 to 35%. The gauges K58 and K59 are only operational from 1991. The station data were checked for duplicate records, anomalies and other issues, and are corrected. Daily rainfall data for the state of Kentucky are available at 22 stations for a period of 46 years (1971–2016). The stations used are numbered as KY# for convenience, and their locations are shown in Figure 3b. The daily rainfall data provided by the Kentucky Agricultural Weather Center, University of Kentucky, do not contain any missing records.

5 | RESULTS AND ANALYSIS

5.1 | Experiment 1: Effect of the size of the neighbourhood

A constant search radius (r_j) of 50 and 300 km were chosen based on the correlogram for Kabini basin and Kentucky state, respectively, for all the experiments. A maximum of 10 stations for Kabini and 8 stations for Kentucky satisfied the r_j criteria at all the stations. Therefore, the number of stations used for interpolation ranged from 2 to 10 and 2 to 8 for Kabini basin and Kentucky, respectively. The error measures were computed with LOOCV, and it was observed from Figure 5 that variant 1 does score better on the metrics for all support set sizes for Kabini basin, but for Kentucky state SM outperforms variant 1 for low station density. The performance of neighbourhood selection has improved as the support set size is increased, although the performance of SM_{NN_d} was better when the support set size is less than 5 for Kentucky state. Furthermore, when the support set size is more than 7, the percentage improvement in error measures was not substantial in particular to the measures CC and d, in both the case studies. Therefore, a support set size of 7 is used for gridded data generation for Kabini basin and the state of Kentucky. The correlation coefficient between the NS and the neighbouring station was computed when both the stations report non-zero rainfall. Furthermore, different correlation measures (Pearson correlation coefficient r, Kendall's tau τ, and Spearman' rho ρ) were used to assess how a specific correlation measure influences the performance of the interpolation.

The parametric method assumes that data are normally distributed, whereas the rank-based method does not require such an assumption. Table 1 shows the error measures with different correlation measures for both the case studies for support set 7. There is no significant difference in the error measures when different types of correlation coefficients are used interpolation in Kabini River basin. However, for Kentucky state, the error measures are marginally better when Spearman's correlation coefficient is used, with the best RMSE among the three. Therefore, for further testing of the variants, only the method based on Spearman's correlation is used.

5.2 | Experiment 2: Sensitivity of length of time series in neighbourhood selection

This experiment was conducted using variant 1 and is compared with SM_{NN_d} using data from the state of Kentucky. This experiment was limited to the region of Kentucky

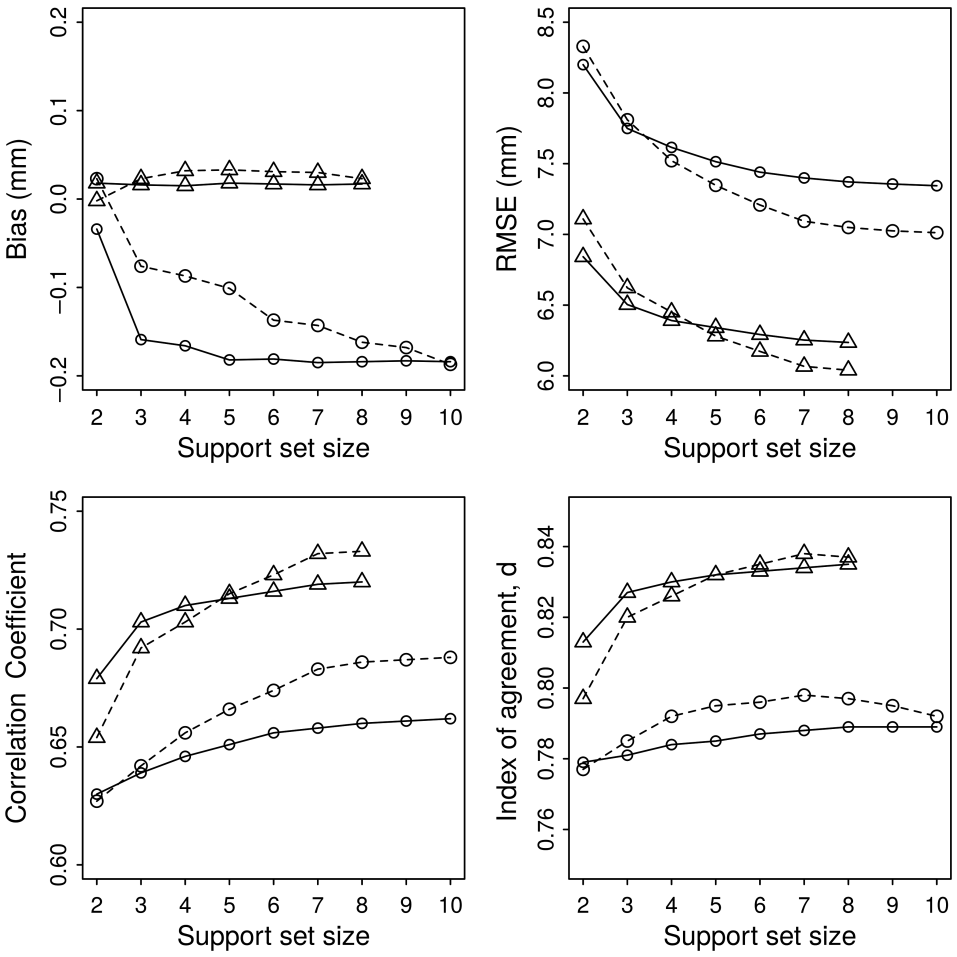

FIGURE 5 Error measures with varying support set size. Solid line with circle marker and solid line with triangle marker is with use of SM for Kabini River basin and Kentucky state, respectively. Similarly, dashed line with circle marker and dashed line with triangle marker is with use of variant 1 for neighbourhood selection

TABLE 1 Variation of error measures with three different measures of correlation for daily rainfall using variant 1 for Kabini basin and Kentucky state

	Case study					
	Kabini			Kentucky		
Error measure	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
Bias (mm)	−0.143	−0.157	−0.159	0.03	0.029	0.022
RMSE (mm)	7.093	7.10	7.165	6.067	6.051	6.154
Correlation	0.683	0.682	0.677	0.732	0.733	0.725
d	0.798	0.797	0.795	0.838	0.838	0.835
NSE	0.46	0.459	0.449	0.53	0.532	0.516

Abbreviations: *d*, index of agreement; NSE, Nash–Sutcliffe model efficiency; RMSE, root-mean-square error.

because of the availability of rainfall data without any gaps. Although the support set is different when the neighbourhood is small, it is mostly identical for each sample as the number of stations is increased. Figure 6 shows the variations in RMSE values for different periods of the time series. The median RMSE across different TL is improving with increase in the support set size and the behaviour is consistent across all TL related experiments. Furthermore, the uncertainty in RMSE decreases with increase in TL. In addition, it is clear from the sensitivity tests, when the number of support stations is more than 6, the variant 1 provides lower RMSE than SM for Kentucky. Similar results were noted not only for the whole (Figure 5) but also for all the shorter lengths of time series.

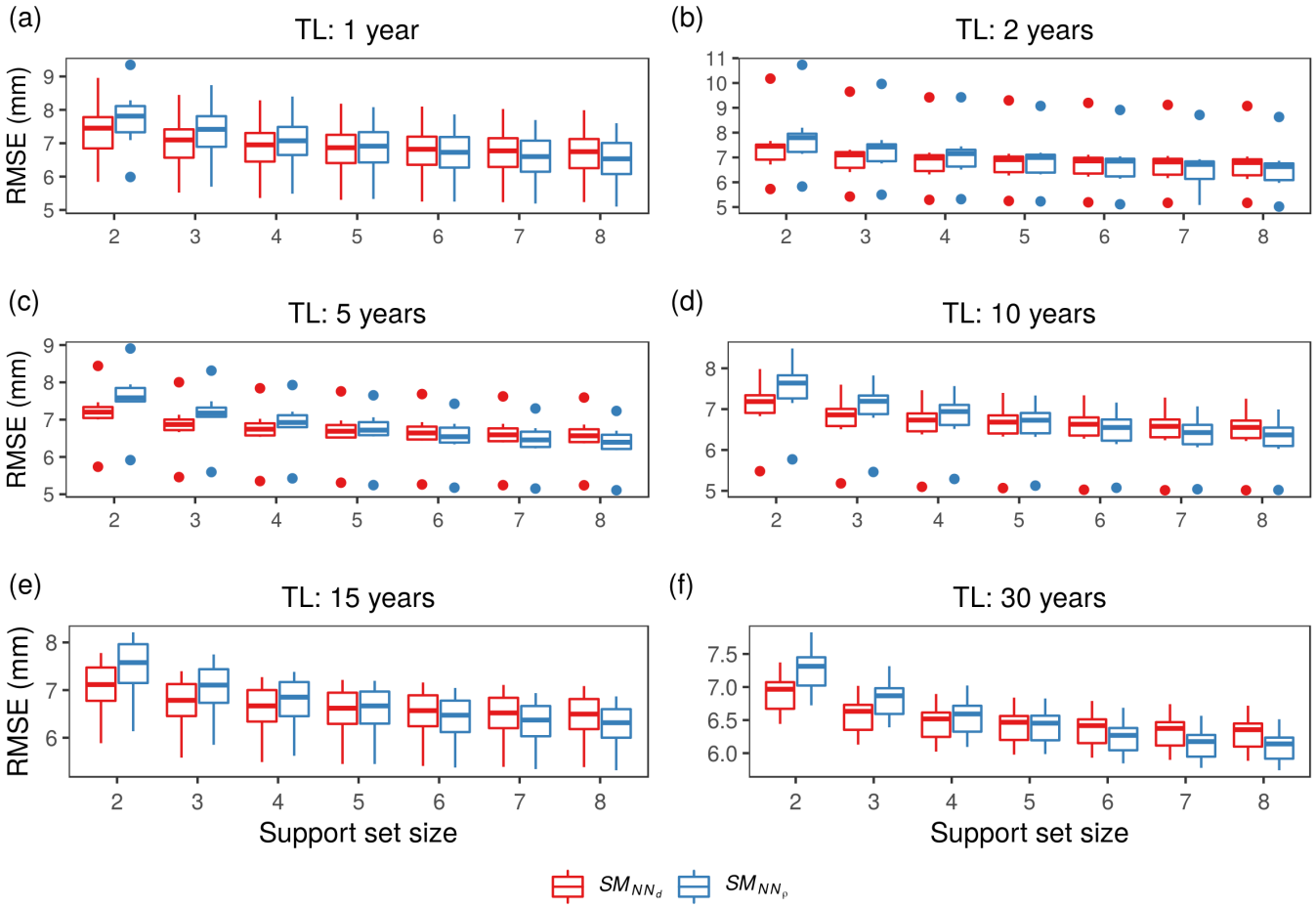

FIGURE 6 Variation of RMSE values with support set size corresponding to different TL (length of time series) [Colour figure can be viewed at wileyonlinelibrary.com]

5.3 | Experiment 3: Effect of choosing the NS in a climatic zone

The daily data from all the stations for only the years with no missing data were used for *k*-means clustering approach. Pascal (1982) and Gunnell and Bourgeon (1997) have identified three zones along the West–East traverse in the Kabini River basin such as humid, semi-humid (transient), and semi-arid based on the steep rainfall gradient. Therefore, we have used three clusters ($k = 3$). In this study continuous data without gaps available for Kabini basin is 10 years and for Kentucky is 46 years. Figure 7a,b shows the three clusters identified by *k*-means clustering approach in both the study regions. In case of missing data imputation or LOOCV, once the clusters are identified, a check is made if NS and BS are from the same cluster, if these two stations (i.e., NS and BS) do not belong to the same cluster, then the next closest station from the distance-sorted list is chosen as NS and the correlations are re-calculated based on the new NS. However, in case of gridded data generation, if the GC of grid is not part of any cluster, then the cluster with the NS determines the neighbourhood selection. Two stations in the Kabini River basin and two stations in Kentucky state are selected, wherein the reference station and the NS are in different cluster groups. In the Kabini basin NS for station K48 in cluster group 1 based on distance is K49, and for K54 in cluster group 2 (Figure 7a) is K52, however with the constraint of cluster group the NS was set to K46 and K49, respectively. Similarly, for the Kentucy state NS based on distance to station KY3 is KY17 and KY8 is KY13 and vice versa. However, based on the clustering, NS for the gauges was corrected to be of the same class as that of gauge, that is, the new NS for KY3 is KY9, and for KY8 it is KY2. After the adjustment to the NS is done, improvement was observed in the error measures. Figure 8 shows the KDEs comparison before and after NS correction with respect to the observed rainfall. The KDEs match well with the observed rainfall after the NS is corrected to be from the same cluster group as the reference station. The same behavior was not significantly visible for Kentucky.

5.4 | Experiment 4: Interpolation using anomalies

The RMSE and correlation values using the anaomalies for support of size 7 for SM and SM_{NN_ρ} are 7.287 and 6.945 mm and 0.676 and 0.701, respectively. There is an improvement in the error metrics when the anaomalies are used rather than absolute values in interpolation. Figure 9 shows the KDEs of the four moments of the observed and

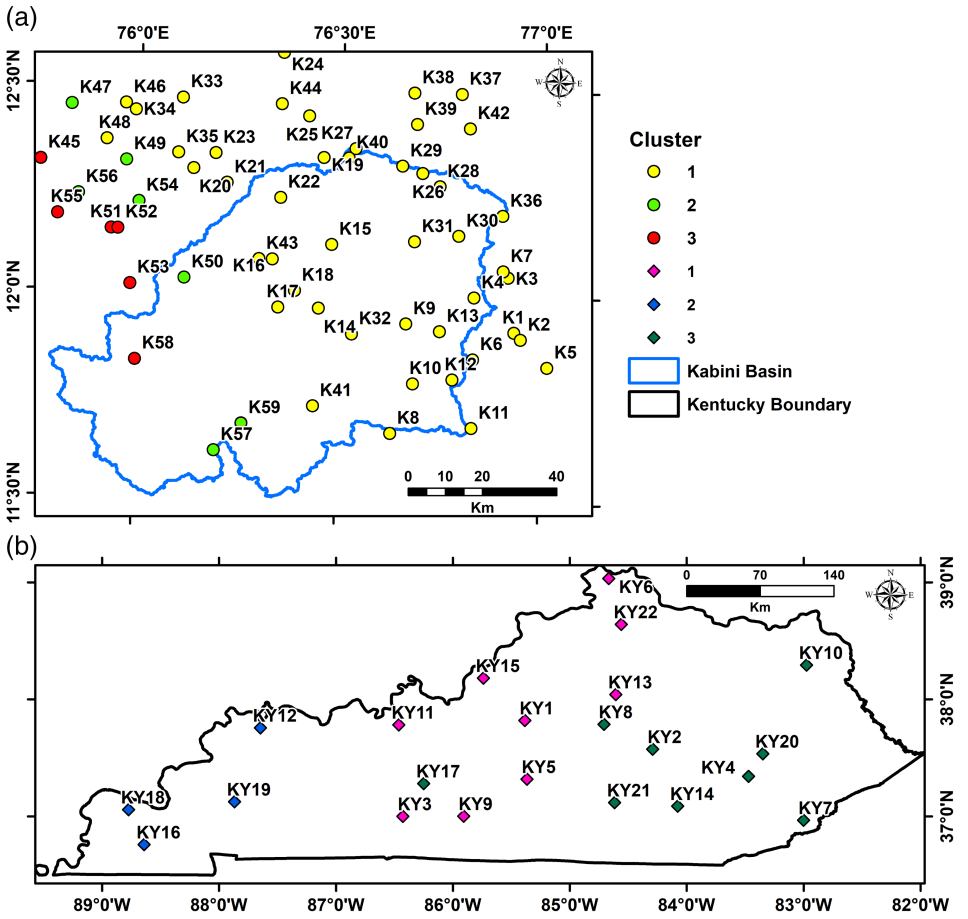

FIGURE 7 Spatial clusters of rainfall gauges defined using *k*-means clustering approach for (a) Kabini River basin and (b) Kentucky state [Colour figure can be viewed at wileyonlinelibrary.com]

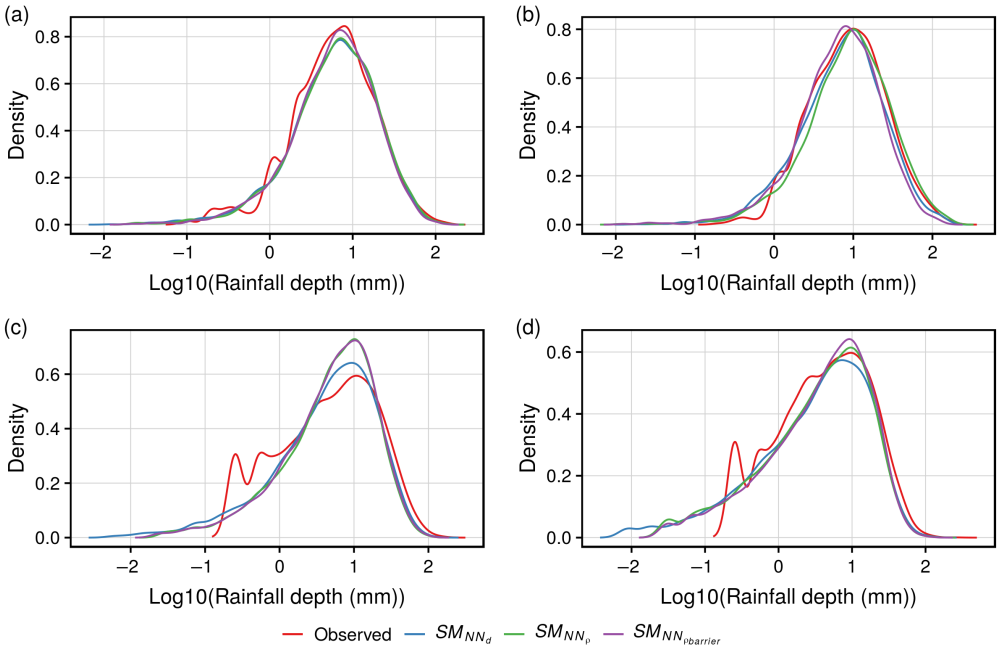

FIGURE 8 KDEs of log10 of observed and interpolated rainfall (a) K48, (b) K54 stations in Kabini River basin and (c) KY3, (d) KY8 stations in Kentucky [Colour figure can be viewed at wileyonlinelibrary.com]

interpolated rainfall for Kabini basin. The first moment is preserved well by both the SM_{NN_d} and with the use of variant 1. The other three moments are overestimated in both the original and the variant 1 and the skewness in the observed data is not preserved.

5.5 | Experiment 5: Neighbourhood selection by distribution similarity

The neighbourhood selection by distribution similarity cannot be applied to daily data because of the high variability rainfall, most of the stations fail the KS test with the NS as reference. The support has a maximum of one or two stations which pass the KS test at the daily scale. Therefore, the variants 2 and 3 experiments are conducted with the monthly data. Inferring from results of Experiments 3 and 4, the CAI approach is adopted, and NS is chosen to be from a predefined cluster or climatic region. Furthermore, the distribution similarity was tested at a seasonal level. The year was divided into three seasons namely pre-monsoon (January–May), monsoon (June–September), and post-monsoon (October–December). Figure 10 shows KDEs of four stations in Kabini where at two locations K9 and K45 the densities match well with the observed rainfall and other two

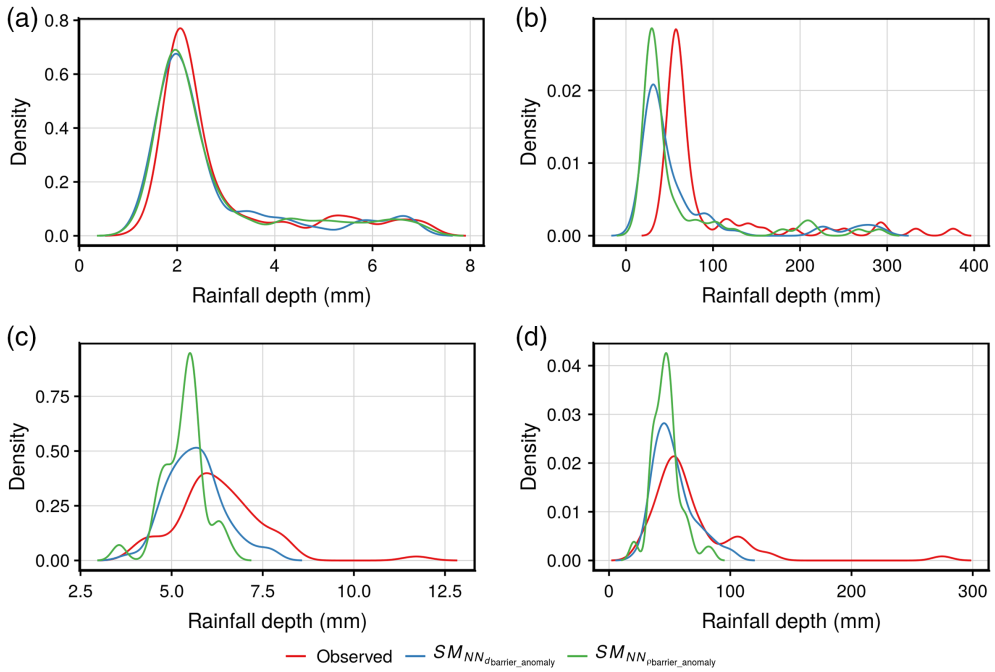

FIGURE 9 KDEs of four moments of observed and interpolated rainfall for Kabini River basin (a) mean, (b) variance, (c) skewness, and (d) kurtosis [Colour figure can be viewed at wileyonlinelibrary.com]

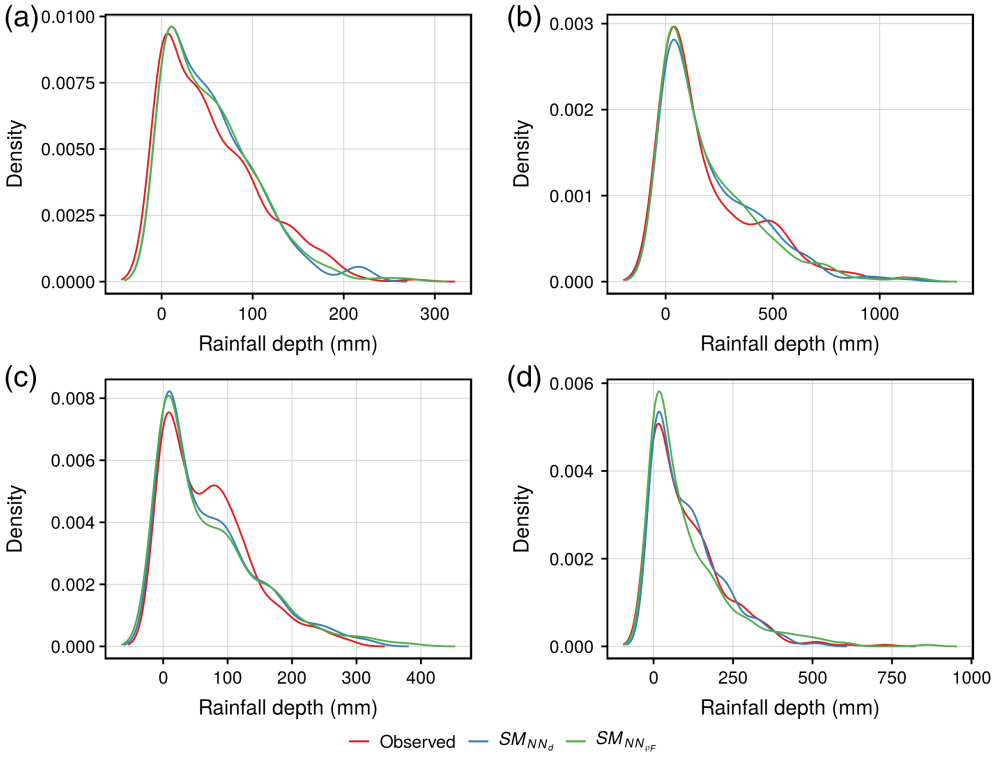

FIGURE 10 KDEs of observed and interpolated monthly rainfall at four locations in the Kabini River basin (a) K9, (b) K45, (c) K16, and (d) K49 [Colour figure can be viewed at wileyonlinelibrary.com]

locations K16 and K49, the densities are overestimated particularly in the right tail of the distribution.

5.6 | Experiment 6: Neighbourhood selection using correlation and distribution similarity

Table 2 shows the error measures and KS test results. Although the error measures are better based on the application of SM_{NN_d} with stations selected by distance, however, with the use of variant 3, it is clear from the KS test results that the number of stations that passed the test are 57 for SM_{NN_d} and 58 for $SM_{NN_{\rho F}}$. The mean of rainfall is well captured by all the three variants, but the other moments are highly left skewed with respect to the observed as shown in Figure 11. In addition, the autocorrelation in the rainfall data at monthly scales needs to be maintained in the interpolated rainfall. To check the persistence, the autocorrelation was computed for four lags. Figure 12 shows the autocorrelation at different temporal lags at three station locations K22, K45, and K57 in the Kabini River basin. It is observed that the autocorrelation is overestimated at few lags in the interpolated data using SM and variant 3. Overestimation of autocorrelation values at several lags in spatially interpolated gridded precipitation data in comparison with rain gauge data was noted in a study by Teegavarapu *et al*. (2017). The use of variant 3 of SM reduced the differences between autocorrelation values based on rain gauge observations and gridded precipitation data. Monthly gridded total rainfall for May 1998 at 5-km spatial resolution is generated using IDW, SM, and variant 1 for Kabini River basin. A constant support set size of 7 is used for all the three interpolations to remove the bias due to varying number of stations. It is evident from Figure 13 that the rainfall variability as reflected in IDW-based interpolated gridded product is preserved better by variant 1 than SM.

TABLE 2 Comparison of the error measures for monthly rainfall in Kabini River basin with SM and its variants 1, 2, and 3 for support set of 5

		Variant		
Error measure	**SM**	**Variant 1**	**Variant 2**	**Variant 3**
ME (mm)	0.493	0.45	0.547	0.504
RMSE (mm)	48.290	46.777	49.609	50.007
Correlation	0.923	0.926	0.919	0.918
d	0.960	0.961	0.958	0.957
Number of sites with confirmed null hypothesis using two-sample KS test	57	57	57	58

5.7 | Experiment 7: Filling missing data

The station KY13 in the Kentucky region was chosen for this experiment. The data were assumed to be missing and was estimated at this station using seven nearby stations.

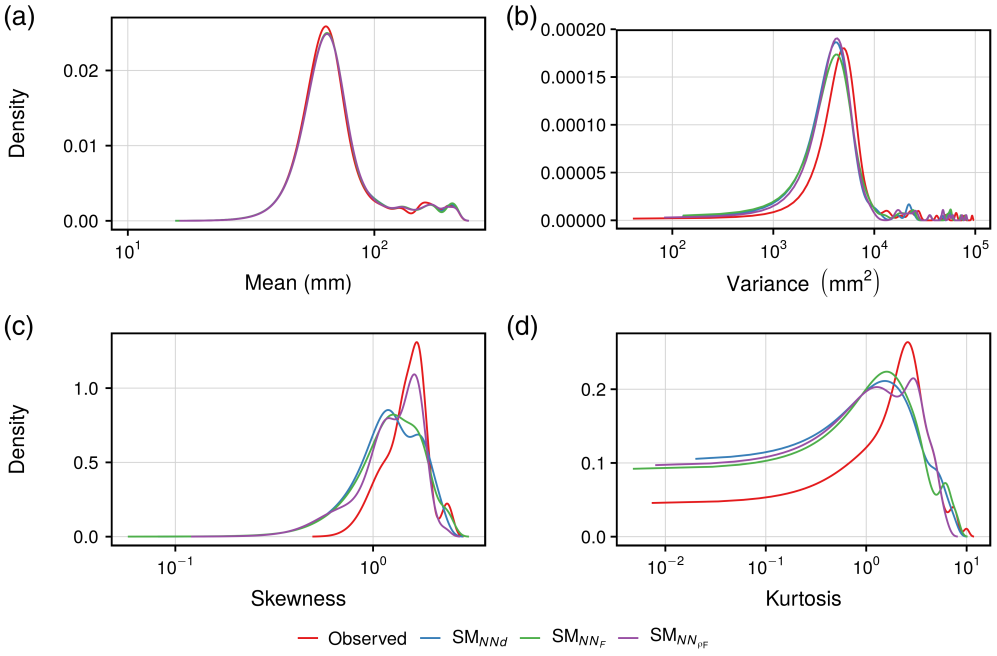

FIGURE 11 KDEs plots of the moments of observed and interpolated monthly rainfall. The *x*-axis is on log10 scale in all panels [Colour figure can be viewed at wileyonlinelibrary.com]

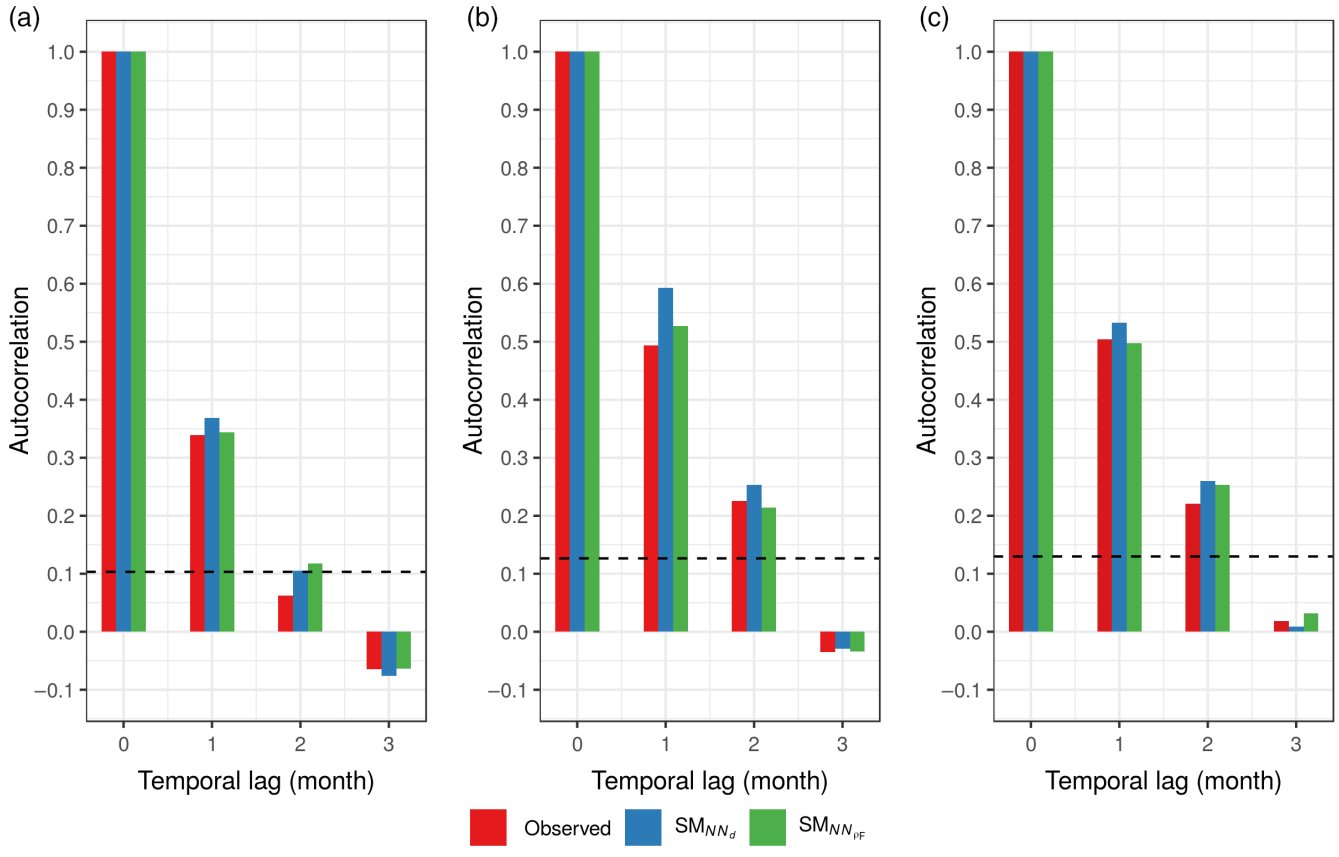

FIGURE 12 Autocorrelation at different lags for the observed and the interpolated rainfall with SM and use of variant 3 for stations (a) K22, (b) K45, and (c) K57. The dotted line corresponds to 95% confidence [Colour figure can be viewed at wileyonlinelibrary.com]

Table 3 provides the error and performance measures from SM and the use of variant 1 for two scenarios. It is apparent based on the measures RMSE, *r*, *d* that variant 1 is better than neighbourhood selected by distance for both the scenarios, with scenario 1 performs better than scenario 2 as the correlations are established with the BS. The control points used for interpolation are same for SM and variant 1 (scenario 2), but are different for scenario 1.

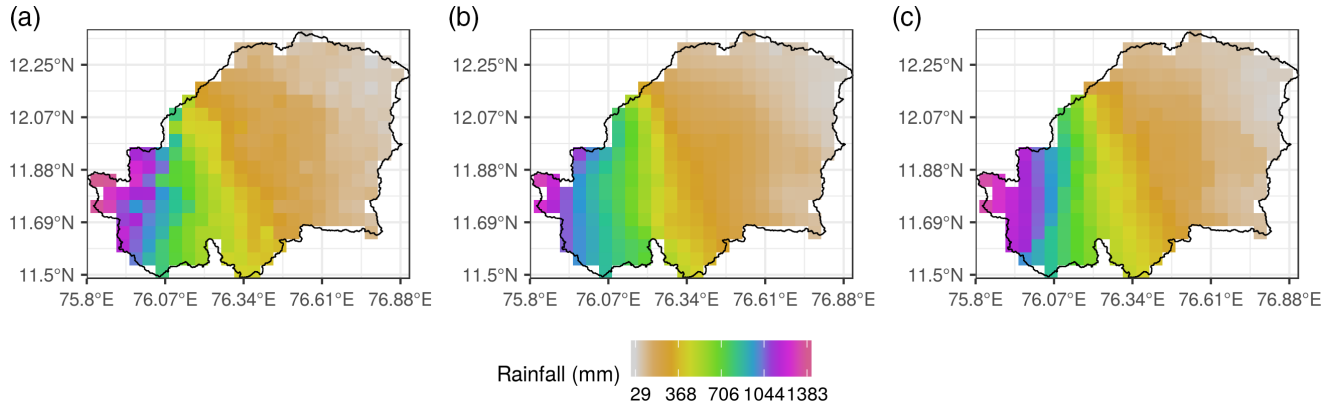

FIGURE 13 Gridded monthly rainfall for May-1998 at 5-km spatial resolution, Kabini River basin using (a) IDW, (b) SM, and (c) variant 1 of SM [Colour figure can be viewed at wileyonlinelibrary.com]

6 | DISCUSSION AND GENERAL REMARKS

The SM that is widely used for generation of gridded rainfall data is re-visited in this study. Conceptual revisions to the neighbourhood selection are proposed and tested for both gridded rainfall data generation and missing data imputation. The results from interpolated estimates were compared with those from the original version of SM. It was found from the experiments conducted that variants of SM work better in regions where there is a gradient in rainfall and are characterized by different climatic zones. The model for gridded data generation was selected from the analysis. Variant 1, SM_{NN_p}, was selected for gridded daily rainfall generation following the CAI approach. The length of the time series is irrelevant for neighbourhood selected by distance; however, for the variants developed in this study it is critical as small samples may not provide robust inference results from the statistical hypothesis test such as KS test conducted in this study. As rainfall characteristics vary over monsoon and non-monsoon months, a seasonal or monthly neighbourhood selection criterion can be developed. Although the distribution similarity criterion has not shown improvement in the error measures, this approach could be tested in regions having a high density of gauges. To compare how the spatial dependence and variability in observations is preserved by the gridded data a variogram is computed using Equation (13). Spatial variance estimates at pre-specific distance intervals based gridded precipitation data were lower than those from the observed point data (figure not shown). The variance deflation observed can be attributed to the main limitation of any spatial interpolation approach which in general underestimates higher-end values thus reducing the variability in estimates. Corrections of gridded precipitation data using quantile mapping or single best estimator (SBE) approaches (Teegavarapu, 2014) or any bias correction method may address the issue of variance deflation. Daily and monthly gridded rainfall data were generated at a resolution of 5×5 km. Changes in the performances of variants for generation of gridded data, w.r.t different spatial resolutions based on a given network density are not evaluated in this study. Future studies may focus on experiments with different grid sizes to arrive at an optimal grid resolution that provides the best estimated values. The success of the variants proposed lies in the appropriate selection of the NS to the centre. If the NS is dissimilar to grid centre rainfall characteristics, then the neighbourhood selected might not result in the best estimate of the gridded rainfall estimate.

TABLE 3 Error measures for different estimation methods used for filling missing data at station KY13 in the Kentucky state

Method	Bias (mm)	RMSE (mm)	Correlation coefficient[c]	*d*	Control points used for interpolation
SM_{NNd}	0.177	7.075	0.508	0.687	KY8, KY2, KY22, KY1, KY15, KY21, and KY5
$SM_{NN\rho}$[a]	0.190	5.460	0.704	0.816	KY1, KY15, KY22, KY8, KY2, KY5, and KY6
$SM_{NN\rho}$[b]	0.247	5.999	0.644	0.781	KY8, KY2, KY5, KY1, KY21, KY22, and KY15

Abbreviations: *d*, index of agreement; NSE, Nash–Sutcliffe model efficiency; RMSE, root-mean-square error.

[a]The correlations are computed between the BS and control points and they are used as weights (scenario 1).

[b]The correlations are computed between the NS and control points and they are used as weights (scenario 2).

[c]Pearson's correlation coefficient.

7 | CONCLUSIONS

Development of daily and monthly gridded precipitation data will significantly advance hydro-climatological studies and also enhance the ability to assess the reliability of satellite-based rainfall products and to evaluate the impacts of climate variability on water resources management. New and improved variants of SM focusing on neighbourhood selection in spatial interpolation for development of gridded precipitation data are investigated in this study. The variants have been tested for their utility in both daily and monthly gridded data generation in two diverse climatological regions. Results suggest that the performance of the variant 1 that uses correlation-based neighbourhood selection was superior for gridded data generation compared to original version of SM at daily scale and monthly scales. The variant 1 could be applied for missing data estimation. The weights for variant 2 that uses distribution similarity criteria for neighbourhood selection are based on distance; however, alternate weighting scheme using the distance from the KS test could be investigated. The proposed variants could also be tested for interpolation of other meteorological variables such as temperature, wind speed, and other hydroclimatological variables as the variants developed are conceptually superior to the distance based selection of neighbourhood. The gridded precipitation data generated using conceptually superior variants evaluated in this study may require corrections for any bias before they are adopted for any hydrologic modelling or climatological studies.

ACKNOWLEDGEMENTS

The authors would like to thank the Department of Economics and Statistics (DES) Karnataka, IMD Thiruvananthapuram and University of Kentucy, United States, for sharing the rainfall data. The first and the third authors would like to acknowledge the financial support received from the Australia-India Strategic Research Fund (DST-AISRF-06220). The second author thanks Fulbright Scholar program for supporting his stay in India during this study. Comments and suggestions from the three reviewers are greatly appreciated.

ORCID

Subash Yeggina https://orcid.org/0000-0002-1390-8275
Ramesh S. V. Teegavarapu https://orcid.org/0000-0002-8194-6038
Sekhar Muddu https://orcid.org/0000-0001-9326-1813

REFERENCES

Ahrens, B. (2006) Distance in spatial interpolation of daily rain gauge data. *Hydrology and Earth System Sciences*, 10, 197–208. https://doi.org/10.5194/hess-10-197-2006.

American Society of Civil Engineers (ASCE). (1996) *Hydrology Handbook*, 2nd edition. Reston, VA: American Society of Civil Engineers (ASCE).

Bárdossy, A. and Li, J. (2008) Geostatistical interpolation using copulas. *Water Resources Research*, 44, W07412. https://doi.org/10.1029/2007WR006115.

Bárdossy, A. (2011) Interpolation of groundwater quality parameters with some values below the detection limit. *Hydrology and Earth System Sciences*, 15, 2763–2775. https://doi.org/10.5194/hess-15-2763-2011.

Burrough, P.A. and McDonnell, R.A. (1998) *Principles of Geographical Information Systems*. Oxford: Oxford University Press.

Buvaneshwari, S., Riotte, J., Sekhar, M., Mohan Kumar, M.S., Sharma, A.K., Duprey, J.-L., Audry, S., Giriraj, P.R., Praveen, Y., Hemanth, M., Durand, P., Braun, J.J. and Ruiz, L. (2017) Groundwater resource vulnerability and spatial variability of nitrate contamination: insights from high density tube well monitoring in a hard rock aquifer. *Science of the Total Environment*, 579, 838–847. https://doi.org/10.1016/j.scitotenv.2016.11.017.

Camera, C., Bruggeman, A., Hadjinicolaou, P., Pashiardis, S. and Lange, M.A. (2014) Evaluation of interpolation techniques for the creation of gridded daily precipitation (1×1 km^2); Cyprus, 1980–2010. *Journal of Geophysical Research: Atmospheres*, 119, 693–712. https://doi.org/10.1002/2013jd020611.

Cressie, N.A.C. (1993) *Statistics for spatial data*. New York, NY: Wiley.

Chen, M., Xie, P., Janowiak, J.E. and Arkin, P.A. (2002) Global land precipitation: a 50-yr monthly analysis based on gauge observations. *Journal of Hydrometeorology*, 3, 249–266. https://doi.org/10.1175/1525-7541(2002)003<0249:glpaym>2.0.co;2.

Chitra-Tarak, R., Ruiz, L., Dattaraja, H.S., Mohan Kumar, M.S., Riotte, J., Suresh, H.S., McMahon, S.M. and Sukumar, R. (2018) The roots of the drought: hydrology and water uptake strategies mediate forest-wide demographic response to precipitation. *Journal of Ecology*, 106, 1495–1507. https://doi.org/10.1111/1365-2745.12925.

Daly, C., Gibson, W.P., Taylor, G.H., Johnson, G.L. and Pasteris, P. (2002) A knowledge-based approach to the statistical mapping of climate. *Climate Research*, 22, 99–113. https://doi.org/10.3354/cr022099.

Eswar, R., Sekhar, M. and Bhattacharya, B.K. (2017a) Comparison of three remote sensing based models for the estimation of latent heat flux over India. *Hydrological Sciences*, 62, 2705–2719. https://doi.org/10.1080/02626667.2017.1404067.

Eswar, R., Sekhar, M., Bhattacharya, B.K. and Bandyopadhyay, S. (2017b) Spatial disaggregation of latent heat flux using contextual models over India. *Remote Sensing*, 9, 949. https://doi.org/10.3390/rs9090949.

Frei, C. and Schär, C. (1998) A precipitation climatology of the Alps from high-resolution rain-gauge observations. *International Journal of Climatology*, 18, 873–900. https://doi.org/10.1002/(sici)1097-0088(19980630)18:8<873::aid-joc255>3.0.co;2-9.

Goovaerts, P. (1997) *Geostatistics for Natural Resources Evaluation*. New York, NY: Oxford University Press.

Gunnell, Y. and Bourgeon, G. (1997) Soils and climatic geomorphology on the Karnataka Plateau, peninsular India. *Catena*, 29, 239–262.

Harris, I., Jones, P.D., Osborn, T.J. and Lister, D.H. (2014) Updated high-resolution grids of monthly climatic observations—the CRU

TS3.10 dataset. *International Journal of Climatology*, 34, 623–642. https://doi.org/10.1002/joc.3711.

Hay, L., Viger, R. and McCabe, G. (1998) Precipitation interpolation in mountainous regions using multiple linear regression. In: *Proceedingsof the HeadWater'98 Conference*. Wallingford: IAHS. IAHS publication number: 248, pp. 33–38.

Hiebl, J. and Frei, C. (2017) Daily precipitation grids for Austria since 1961 development and evaluation of a spatial dataset for hydroclimatic monitoring and modelling. *Theoretical and Applied Climatology*, 132, 327–345. https://doi.org/10.1007/s00704-017-2093-x.

Hofstra, N., Haylock, M., New, M. and Jones, P.D. (2009) Testing E-OBS European high-resolution gridded data set of daily precipitation and surface temperature. Calculation of gridded precipitation data for the global land-surface using in-situ gauge observations. *Journal of Geophysical Research*, 114, D21101. https://doi.org/10.1029/2009jd011799.

Hofstra, N. and New, M. (2009) Spatial variability in correlation decay distance and influence on angular-distance weighting interpolation of daily precipitation over Europe. *International Journal of Climatology*, 29, 1872–1880. https://doi.org/10.1002/joc.1819.

Hwang, Y., Clark, M., Rajagopalan, B. and Leavesley, G. (2012) Spatial interpolation schemes of daily precipitation for hydrologic modeling. *Stochastic Environmental Research and Risk Assessment*, 26, 295–320. https://doi.org/10.1007/s00477-011-0509-1.

Hunter, R.D. and Meentemeyer, R.K. (2005) Climatologically aided mapping of daily precipitation and temperature. *Journal of Applied Meteorology*, 44, 1501–1510. https://doi.org/10.1175/jam2295.1.

Hutchinson, M.F. (1995) Interpolating mean rainfall using thin plate smoothing splines. *International Journal of Geographical Information Science*, 9, 385–403. https://doi.org/10.1080/02693799508902045.

Isotta, F.A., Frei, C., Weilguni, V., Tadić, M.P., Lassègues, P., Rudolf, B., Pavan, V., Cacciamani, C., Antolini, G., Ratto, S.M., Munari, M., Micheletti, S., Bonati, V., Lussana, C., Ronchi, C., Panettieri, E., Marigo, G. and Vertačnik, G. (2014) The climate of daily precipitation in the Alps: development and analysis of a high-resolution grid dataset from pan-Alpine rain-gauge data. *International Journal of Climatology*, 34, 1657–1675. https://doi.org/10.1002/joc.3794.

Kleber, J.E. (Ed.). (2015) *The Kentucky Encyclopedia*. Lexington, KY: University Press of Kentucky.

Kottek, M., Grieser, J., Beck, C., Rudolf, B. and Rubel, F. (2006) World map of the Köppen–Geiger climate classification updated. *Meteorologische Zeitschrift*, 15, 259–263. https://doi.org/10.1127/0941-2948/2006/0130.

Li, J. and Heap, A.D. (2011) A review of comparative studies of spatial interpolation methods in environmental sciences: performance and impact factors. *Ecological Informatics*, 6, 228–241. https://doi.org/10.1016/j.ecoinf.2010.12.003.

Lin, G.-F. and Chen, L.-H. (2004) A spatial interpolation method based on radial basis function networks incorporating a semivariogram model. *Journal of Hydrology*, 288, 288–298. https://doi.org/10.1016/j.jhydrol.2003.10.008.

Livneh, B., Rosenberg, E.A., Lin, C., Nijssen, B., Mishra, V., Andreadis, K.M., Maurer, E.P. and Lettenmaier, D.P. (2013) A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States: update and extensions. *Journal of Climate*, 26, 9384–9392. https://doi.org/10.1175/jcli-d-12-00508.1.

Ly, S., Charles, C. and Degré, A. (2011) Geostatistical interpolation of daily rainfall at catchment scale: the use of several variogram models in the Ourthe and Ambleve catchments Belgium. *Hydrology and Earth System Sciences*, 15, 2259–2274. https://doi.org/10.5194/hess-15-2259-2011.

Malhi, Y. and Wright, J. (2004) Spatial patterns and recent trends in the climate of tropical rainforest regions. *Philosophical Transactions of the Royal Society: B*, 359, 311–329. https://doi.org/10.1098/rstb.2003.1433.

Mitchell, T.D. and Jones, P.D. (2005) An improved method of constructing a database of monthly climate observations and associated high-resolution grids. *International Journal of Climatology*, 25, 693–712. https://doi.org/10.1002/joc.1181.

New, M., Hulme, M. and Jones, P. (2000) Representing twentieth-century space–time climate variability. Part II: development of 1901–96 monthly grids of terrestrial surface climate. *Journal of Climate*, 13, 2217–2238. https://doi.org/10.1175/1520-0442(2000)013<2217:RTCSTC>2.0.CO;2.

Nguyen, X.T., Ngo, D.T., Kamimera, H., Trinh, T.L., Matsumoto, J., Inoue, T. and Phan, V.T. (2016) The Vietnam gridded precipitation (VnGP) dataset: construction and validation. *SOLA*, 12, 291–296. https://doi.org/10.2151/sola.2016-057.

Nychka, D., Furrer, R. and Paige, J. and Sain, S. (2017) *fields: tools for spatial data.* R package version 9.6. doi:10.5065/D6W957CT.

Pai, D.S., Sridhar, L., Rajeevan, M., Sreejith, O.P., Satbhai, N.S. and Mukhopadhyay, B. (2014) Development of a new high spatial resolution (0.25 × 0.25) long period (1901–2010) daily gridded rainfall data set over India and its comparison with existing data sets over the region. *Mausam*, 65, 1–18.

Pascal, J.P. (1982) *Bioclimates of the Western Ghats*. Pondicherry: French Institute of Pondicherry.

Perry, M. and Hollis, D. (2005) The generation of monthly gridded datasets for a range of climatic variables over the UK. *International Journal of Climatology*, 25, 1041–1054. https://doi.org/10.1002/joc.1161.

Rajeevan, M., Bhate, J., Kale, J.D. and Lal, B. (2005) *Development of a High Resolution Daily Gridded Rainfall Data for the Indian Region*. *Met Monograph Climatology 22/2005*. Pune: India Meteorological Department, 26 pp.

Ramos-Calzado, P., Gómez-Camacho, J., Pérez-Bernal, F. and Pita-López, M.F. (2008) A novel approach to precipitation series completion in climatological datasets: application to Andalusia. *International Journal of Climatology*, 28, 1525–1534. https://doi.org/10.1002/joc.1657.

Rudolf, B. and Schneider, U. (2005) Calculation of gridded precipitation data for the global land-surface using in-situ gauge observations. In: *Proceedings of the Second Workshop of the International Precipitation Working Group* IPWG, October 2004, Monterey, CA. Darmstadt: EUMETSAT, pp. 231–247.

Ruelland, D., Ardoin-Bardin, S., Billen, G. and Servat, E. (2008) Sensitivity of a lumped and semi-distributed hydrological model to several methods of rainfall interpolation on a large basin in West Africa. *Journal of Hydrology*, 361, 96–117. https://doi.org/10.1016/j.jhydrol.2008.07.049.

Sekhar, M., Riotte, J., Ruiz, L., Jouquet, J. and Braun, J.J. (2016) Influences of climate and agriculture on water and biogeochemical cycles: Kabini Critical Zone Observatory. *Proceedings of the Indian National Science Academy*, 82, 833–846. https://doi.org/10.16943/ptinsa/2016/48488.

Sekhon, J.S. (2011) Multivariate and propensity score matching software with automated balance optimization: the matching package for R. *Journal of Statistical Software*, 42, 1–52. https://doi.org/10.18637/jss.v042.i07.

Sharma, A., Hubert-Moy, L., Buvaneshwari, S., Sekhar, M., Ruiz, L., Bandyopadhyay, S. and Corgne, S. (2018) Irrigation history estimation using multitemporal Landsat satellite images: application to an intensive groundwater irrigated agricultural watershed in India. *Remote Sensing*, 10, 893. https://doi.org/10.3390/rs10060893.

Shepard, D. (1968) A two dimensional interpolation function for regularly spaced data. In: *Proceedings of the 23d National Conference of the Association for Computing Machinery*. Princeton, NJ: ACM, pp. 517–524. https://doi.org/10.1145/800186.810616.

Shepard, D.S. (1984) Computer mapping: the SYMAP interpolation algorithm. In: Gaile, G.L. and Willmott, C.J. (Eds.) *Spatial Statistics and Models*. Dordrecht: D. Reidel, pp. 133–145.

Simolo, C., Brunetti, M., Maugeri, M. and Nanni, T. (2010) Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach. *International Journal of Climatology*, 30, 1564–1576. https://doi.org/10.1002/joc.1992.

Sreelash, K., Buis, S., Sekhar, M., Ruiz, L., Tomer, S.K. and Guerif, M. (2017) Estimation of available water capacity components of two-layered soils using crop model inversion: effect of crop type and water regime. *Journal of Hydrology*, 546, 166–178. https://doi.org/10.1016/j.jhydrol.2016.12.049.

Srivastava, A.K., Rajeevan, M. and Kshirsagar, S.R. (2009) Development of a high resolution daily gridded temperature data set (1969–2005) for the Indian region. *Atmospheric Science Letters*, 10, 249–254. https://doi.org/10.1002/asl.232.

Subash, Y., Sekhar, M., Tomer, S.K. and Sharma, A.K. (2017) A framework for assessment of climate change impacts on the groundwater system. In: Ojha, C.S.P. , Rao, S. , Zhang, T. and Bardossy, A. (Eds.) *Sustainable Water Resources Management*, Vol. 14. Reston, VA: ASCE, pp. 375–397. https://doi.org/10.1061/9780784414767.ch14.

Teegavarapu, R.S.V. and Chandramouli, V. (2005) Improved weighting methods deterministic and stochastic data-driven models for estimation of missing precipitation records. *Journal of Hydrology*, 312, 191–206. https://doi.org/10.1016/j.jhydrol.2005.02.015.

Teegavarapu, R.S.V., Goly, A. and Wu, Q. (2017) Comprehensive framework for assessment of radar-based precipitation data estimates. *Journal of Hydrologic Engineering*, 22, E4015002. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001277.

Teegavarapu, R.S.V. (2007) Use of universal function approximation in variance-dependent surface interpolation method: an application in hydrology. *Journal of Hydrology*, 332, 16–29. https://doi.org/10.1016/j.jhydrol.2006.06.017.

Teegavarapu, R.S.V. (2014) Statistical corrections of spatially interpolated precipitation estimates. *Hydrological Processes*, 28, 3789–3808. https://doi.org/10.1002/hyp.9906.

Tomer, S.K., Al Bitar, A., Sekhar, M., Zribi, M., Bandyopadhyay, S., Sreelash, K., Sharma, A.K., Corgne, S. and Kerr, Y. (2015) Retrieval and multi-scale validation of soil moisture from multi-temporal SAR data in a semi-arid tropical region. *Remote Sensing*, 7, 8128–8153. https://doi.org/10.3390/rs70608128.

Wasko, C., Sharma, A. and Rasmussen, P. (2013) Improved spatial prediction: a combinatorial approach. *Water Resources Research*, 49, 3927–3935. https://doi.org/10.1002/wrcr.20290.

Wichern, D.W. and Johnson, R.A. (1992) *Applied Multivariate Statistical Analysis*. Englewood Cliffs, NJ: Prentice Hall.

Wilcox, R. (2005) Kolmogorov–Smirnov test. In: Armitageand, P. and Colton, T. (Eds.) *Encyclopedia of Biostatistics*. NewYork, NY: Wiley. https://doi.org/10.1002/0470011815.b2a15064.

Willmott, C.J. (1981) On the validation of models. *Physical Geography*, 2, 184–194.

Willmott, C.J., Rowe, C.M. and Philpot, W.D. (1985) Small-scale climate maps: a sensitivity analysis of some common assumptions associated with grid-point interpolation and contouring. *American Cartographer*, 12, 5–16. https://doi.org/10.1559/152304085783914686.

Willmott, C.J. and Robeson, S.M. (1995) Climatologically aided interpolation (CAI) of terrestrial air temperature. *International Journal of Climatology*, 15, 221–229. https://doi.org/10.1002/joc.3370150207.

Xavier, A.C., King, C.W. and Scanlon, B.R. (2015) Daily gridded meteorological variables in Brazil (1980–2013). *International Journal of Climatology*, 36, 2644–2659. https://doi.org/10.1002/joc.4518.

Yanto, Livneh, B. and Rajagopalan, B. (2017) Development of a gridded meteorological dataset over Java island Indonesia 1985–2014. *Scientific Data*, 4, 170072. https://doi.org/10.1038/sdata.2017.72.

Yatagai, A., Kamiguchi, K., Arakawa, O., Hamada, A., Yasutomi, N. and Kitoh, A. (2012) APHRODITE: constructing a long-term daily gridded precipitation dataset for Asia based on a dense network of rain gauges. *Bulletin of the American Meteorological Society*, 93, 1401–1415. https://doi.org/10.1175/bams-d-11-0012.

How to cite this article: Yeggina S, Teegavarapu RSV, Muddu S. A conceptually superior variant of Shepard's method with modified neighbourhood selection for precipitation interpolation. *Int J Climatol*. 2019;39:4627–4647. https://doi.org/10.1002/joc.6091