Hindawi Publishing Corporation
Advances in Meteorology
Volume 2015, Article ID 935868, 12 pages
http://dx.doi.org/10.1155/2015/935868

Research Article

Interpolation of Missing Precipitation Data Using Kernel Estimations for Hydrologic Modeling

Hyojin Lee[1] and Kwangmin Kang[2]

[1] *APEC Climate Center, 12 Centum 7-ro, Haeundae-gu, Busan 612-020, Republic of Korea*
[2] *School of Agriculture and Natural Science, University of Maryland, College Park, MD 20742, USA*

Correspondence should be addressed to Kwangmin Kang; hbkangkm@gmail.com

Received 17 October 2014; Revised 15 January 2015; Accepted 5 February 2015

Academic Editor: Fubao Sun

Precipitation is the main factor that drives hydrologic modeling; therefore, missing precipitation data can cause malfunctions in hydrologic modeling. Although interpolation of missing precipitation data is recognized as an important research topic, only a few methods follow a regression approach. In this study, daily precipitation data were interpolated using five different kernel functions, namely, Epanechnikov, Quartic, Triweight, Tricube, and Cosine, to estimate missing precipitation data. This study also presents an assessment that compares estimation of missing precipitation data through Kth nearest neighborhood (KNN) regression to the five different kernel estimations and their performance in simulating streamflow using the Soil Water Assessment Tool (SWAT) hydrologic model. The results show that the kernel approaches provide higher quality interpolation of precipitation data compared with the KNN regression approach, in terms of both statistical data assessment and hydrologic modeling performance.

1. Introduction

Precipitation data are key factors in hydrologic modeling for estimating rainfall-runoff mechanism [1]. Malfunctions in running hydrologic modeling can occur due to non-continuous time series precipitation inputs. In light of this important issue, estimation of missing precipitation data is a challenging task for hydrologic modeling. Many hydrologic modeling require interpolation of missing precipitation data [2], meteorological data series completion [3], or imputation of meteorological data [4]. To estimate missing precipitation, researchers should consider spatiotemporal variations in precipitation (rainfall and snowfall) values and the related physical processes. However, accounting for spatial-temporal variation and physical processes can be difficult if there is a lack of equipment for measuring precipitation. Thus, statistical approaches have emerged as widely used methods for filling in missing precipitation data [5].

Many studies have investigated supplanting missing streamflow data with several statistical approaches [5], but there are limited studies on the interpolation of incomplete precipitation and temperature data [6–10]. Recently, the investigation of artificial neural networks (ANNs: [11]), a more advanced statistical approach, to estimate missing precipitation data, has been proposed [12]. ANNs can learn from training data to reconstruct a nonlinear relationship and obtain values for missing data. Pisoni et al. [13] investigated the interpolation of missing data for sea surface temperature (SST) satellite images using the ANN method; they found that the results from the ANN approach show better accuracy than the results from an interpolation system, as suggested by Seze and Desbois (1987). Nevertheless, ANNs are still under dispute because their neuron systems cannot provide clear relationships between data [14].

The American Society of Civil Engineers (ASCE) Task Committee [15] discussed that although the performance of ANNs for estimating missing precipitation data has already been verified, an alternate solution should be suggested for cases in which the available data are insufficient due to the reliance of ANNs on high data quality and quantity. Additionally, ANNs have other limitations, such as a lack of physical concepts and relations, based on the experience and preferences of those using, studying, and training the networks [15–17]. Since ANNs are regarded as black-box model [18], it is difficult to use this method for realizing more linear relationships, even though ANNs can achieve convergence for almost any problem [17]. Thus, for real mechanisms in hydrologic models, in which linear relationships exist between series of weather inputs, the solution is less explicit [19].

Generally, a regression or a distance weighted method is most commonly used for estimating missing precipitation for hydrologic modeling [20]. Daly et al. [21] also propose a variety of regression models to incorporate spatial variation in weather data. However, Creutin et al. [22] found that even though simple linear regression of interpolation approaches show satisfactory serial correlation of daily or monthly streamflow; precipitation patterns do not show proper correlation when simple linear regression or interpolation approaches are used. Furthermore, if a regression method is used for estimating missing precipitation to make refined precipitation time series, a small data sample would not follow the normal distribution based on basic theory of linear regression.

Another approach for estimating missing precipitation data to use neighboring data is based on distance weight. Xia et al. [23] used the closest station to reconstruct missing precipitation data through geometrical distance weight; Willmott et al. [24] used arithmetic data averaging from neighboring data to filling missing precipitation; and Teegavarapu and Chandramouli [25] used an inverse distance weight method from neighboring data to estimate missing precipitation data. Smith [26], Simanton and Osborn [27], and Salas [28] suggest that traditional weighting and data-driven methods, namely, distance based weighting methods, are interpolated for estimating missing precipitation data. Distance weight approaches for estimating missing precipitation data are combined with linear regression and median distribution of regression [29, 30]. Young [31] and Filippini et al. [32] suggested spatially interpolating the correlation to define weight in terms of each station.

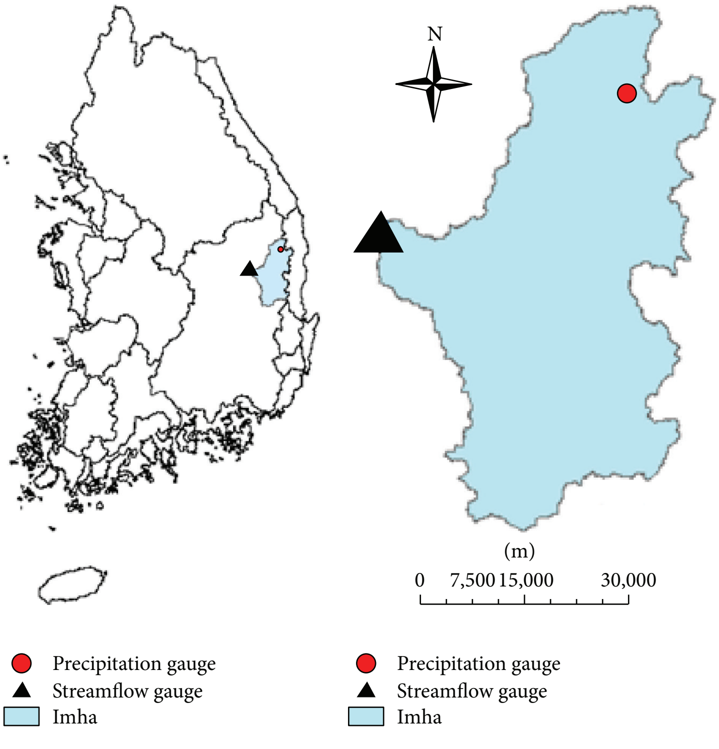
Estimation of missing precipitation data is possible when data are available for the same location. Linacre (1992) investigated the interpolation of missing precipitation data by using the mean value of a data series at the same location and Lowry [33] suggested simple interpolation between available data series. Acock and Pachepsky [34] used data from several days before and after missing precipitation data points for estimating the incomplete precipitation data. K-nearest neighborhood ($k$nn) regression is a basic method for estimating missing precipitation data that considers vicinity. However, the method has some weaknesses when the data have outliers or a nonlinear trend exists around the missing data. While $k$nn regression has a fundamental assumption to follow a normal distribution which is statistically unsound, the kernel method uses a mean value, which can overcome $k$nn regression's weakness through the kernel weighting method. By using neighbor data in a kernel function, even though the data show a nonlinear trend, it can overcome $k$nn regression weakness.

The objective of this study was to reconstruct daily precipitation data by using five different kernel functions (Epanechnikov, Quartic, Triweight, Tricube, and Cosine) to estimate missing precipitation data. This study also presents an assessment that compares estimation of missing precipitation data through $k$nn regression to the five different kernel estimations and their performance in simulating streamflow using the Soil Water Assessment Tool (SWAT) hydrologic model. The remainder of this paper is organized as follows. Section 2 provides a description of the study area and the hydrologic model. In Section 3, the methodology of the five different kernel methods is presented. Section 4 presents the results of the interpolation of the missing daily precipitation data and the hydrologic model simulation. Finally, conclusions are in Section 5.

2. Study Area and Hydrologic Model

The Imha (Figure 1) watershed was selected as the test bed for this study. The Imha watershed is a tributary of the Nakdong River basin and is located in the upper side of the Nakdong River basin in South Korea. It is characterized by a mountainous area; approximately 79.8% of the total area of 1,361 km^2 is mountainous. The slope in the Imha watershed is 40% to 60%, that is, 655 km^2 as 33% of total watershed area. The elevation of the Imha watershed ranges from 80 to 1215 m. The average annual precipitation, minimum temperature, maximum temperature, humidity, and wind speed for the Imha watershed are 1,050 mm, 7°C, 18.8°C, 65%, and 1.6 m/s, respectively (Water Management Information System (WAMIS), http://www.wamis.go.kr/). Since the climate conditions in this area are defined by warm temperatures, there is no precipitation in the form of snow; all precipitation consists of rainfall. For this evaluation of interpolation of precipitation data and hydrologic model performance, precipitation and streamflow gauges were selected as shown in Figure 1 and precipitation and streamflow data were sourced from the Water Management Information System (http://www.wamis.go.kr/).

This study selected the SWAT model for analysis. SWAT has a GIS extension, ArcSWAT, which allows the use of various GIS based datasets to model the geomorphology of a given basin. The SWAT model was developed through research by the USDA (United States Department of Agriculture), Agricultural Research Service (ARS). Major data inputs for SWAT include temperature (maximum and minimum), daily precipitation, solar radiation, relative humidity, wind speed, and geospatial data representing soil types, land cover, and elevation. A watershed is divided into smaller subbasins, which must be broken up into smaller units known as hydrologic response units (HRU). Each of these HRUs is characterized by uniform land use and soil type. SWAT can be used to accurately predict hydrologic patterns for extended periods of time [35]. Canopy interception is implicit in the curve number (CN) method and is explicit for the Green-Ampt method. Infiltration is most accurately accounted for using the CN method in SWAT. An alternative method may be used to account for infiltration is the Green-Ampt method. However, the Green-Ampt method has not been shown to increase accuracy over the CN method, thus the CN method was used in this study.



(a)

(b)

Figure 1: Study basin locations including rain and stream gauges (left figure: map of South Korea; right figure: Imha watershed).

3. Methodology

This study used the five kernel functions, Epanechnikov, Quartic, Triweight, Tricube, and Cosine, as a weight to predict missing values. Tricube method has large weight around target point. Even though Tricube weight is similar to Triweight, the decreasing acceleration of weight as far away from target point is less than Triweight. Next higher weight around target point is Quartic, which speed in decreasing weight is similar to Triweight. Both Epanechnikov and Cosine have small effect on neighboring values. A brief description of the five kernel functions and their application for reconstructing the missing values is presented in the following and specific kernel functions are described in Appendix A.

3.1. Epanechnikov. The Epanechnikov kernel is the most often used kernel function. The Epanechnikov kernel assigns zero weight to observations that are a distance of four, six, and eight away from the reference point. These values correspond to the choice of the interval width. This is often called the choice of smoothing parameter or band width selection. The main character of the Epanechnikov kernel is that even though the distance is far away from target value, namely, the missing value in this research, its estimation is smooth. A brief description is given by the following:

$$K(x) = \frac{3}{4}\left(1 - x^2\right), \tag{1}$$

where $K(x)$ is the kernel function and x is surrounding the nearest value as an independent in data.

3.2. Quartic. The second kernel function used in this research was the Quartic kernel which has more weight sensitivity based on distance from the missing value. Since the applied weight is largely different between near and far data points, it is more influenced by surrounding data. It consists of a fourth-order equation which has more sensitivity in terms of distance than second-order equation. It is described by the following:

$$K(x) = \frac{15}{16}\left(1 - x^2\right)^2. \tag{2}$$

3.3. Triweight. The third kernel function used in this research was the Triweight kernel which consists of a sixth-order equation. It has the most sensitivity in terms of distance because a sixth-order equation estimates the missing value based on the difference in distance with a weighted function as shown by the following:

$$K(x) = \frac{35}{32}\left(1 - x^2\right)^3. \tag{3}$$

3.4. Tricube. The fourth kernel function used in this research was the Tricube kernel, which uses absolute values. Since it uses absolute values, it presents a smoother pattern for nearest values than the Triweight kernel. However, as

Table 1: Results of normality test with Shapiro-Wilk method for each K-nearest neighborhood. DF represents degree of freedom and P value means significance probability.

	4-NN			6-NN			8-NN		
	W	DF	P value	W	DF	P value	W	DF	P value
Ep	0.808	19	0.0015	0.740	19	0.0002	0.766	19	0.0004
Qu	0.831	19	0.0033	0.768	19	0.0004	0.721	19	0.0001
Tw	0.827	19	0.0029	0.789	19	0.0008	0.745	19	0.0002
Tc	0.839	19	0.0045	0.764	19	0.0004	0.742	19	0.0002
Co	0.817	19	0.0020	0.742	19	0.0002	0.763	19	0.0003
Reg	0.876	19	0.0186	0.858	19	0.0089	0.883	19	0.0242

(Ep: Epanechnikov, Qu: Quartic, Tw: Triweight, Tc: Tricube, Co: Cosine, and Reg: regression).

the values move further away from the nearest values, it shows a steep trend. The Tricube kernel has the most sensitivity in terms of weighted distance due to the fact that it consists of a ninth-order equation, as shown in the following:

$$K(x) = \frac{70}{81}\left(1 - |x|^3\right)^3. \tag{4}$$

3.5. Cosine. The fifth kernel function used in this research was the Cosine kernel function. It is a widely applied kernel function in various fields because it has a constant curvature. Its shape is similar to the Epanechnikov kernel, even though it uses a cosine function as shown in the following:

$$K(x) = \frac{\pi}{4}\cos\left(\frac{\pi}{2}x\right). \tag{5}$$

3.6. Calculation of the Missing Value. After using a kernel function to calculate the weight of the missing data, estimation of the missing data is performed using the following:

$$\begin{gathered} M = \frac{1}{P}\sum_{i=1}^{P} x_i \cdot K(u_i), \\ u_i = \frac{N_i}{0.5P + 1}, \\ N_i = -\frac{P}{2}, \dots, \frac{P}{2}, \end{gathered} \tag{6}$$

where M is the missing value, P is the number of the nearest neighborhood, and u_i is the Nth nearest values which correspond to x_i (positive means the right side and negative means the left side). The kernel function should have bilateral symmetry based on a value of zero. If using, for example, the four nearest neighborhoods for estimating the missing value, the neighborhood values used will be two from right side and another two from left side. The specific equation for this example is shown in the following and example calculation is described in Appendix B:

$$\begin{aligned} M = \frac{1}{4}\Bigl\{ & K\left(-\frac{2}{3}\right)\cdot x_1 + K\left(-\frac{1}{3}\right)\cdot x_2 \\ & + K\left(\frac{1}{3}\right)\cdot x_3 + K\left(\frac{2}{3}\right)\cdot x_4 \Bigr\}. \end{aligned} \tag{7}$$

3.7. Statistic Tests. A normality test is required to evaluate for infilling the methods for filling in interpolation data. The Shapiro-Wilk [36] normality test was used with nineteen samples to determine whether the average difference is normally distributed or not. The test statistic is as shown in the following:

$$W = \frac{\left(\sum_{i=1}^{n} a_i y_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}, \tag{8}$$

where y_i is the ith order statistic, namely, the ith smallest value in the sample, $\overline{y}$ is the mean of y_i, and a_i is a constant given by ordered data. The null hypothesis of the Shapiro-Wilk normality test is that sample is normally distributed, and if significance probability is less than 5%, the null hypothesis will be denied, meaning the sample does not satisfy normal distribution. Since the significance probability for the entire group (Table 1) is below 5%, the null hypothesis is denied. This study should, therefore, use a nonparametric test for normality analysis.

The Friedman test [37], which is a kind of k-sample test that can provide the difference between paired values, was selected as a nonparametric test. This method evaluates a small sample for differences by ranking a sequence list. The null hypothesis of the Friedman test is that there is no average difference in each group and if the significance probability is less than 5%, the null hypothesis will be denied, thus conducting that in each group exists an average difference. A brief description of Friedman test is in the following:

$$Q = \frac{\mathrm{SS_t}}{\mathrm{SS_e}}, \tag{9}$$

where $\mathrm{SS_t}$ and $\mathrm{SS_e}$ are the sum of the squared treatment and sum of the squared error, respectively.

The null hypothesis in this instance was denied because the significance probability was less than 5% for each and this study concluded that each interpolation method has an average difference, which is why each method is considered independent, even though this study used five different kernel methods. For example, the average rank for four reference points for $k$nn-regression, Tricube, Quartic, Cosine, Triweight, and Epanechnikov varies from a large average to a small average rank (Table 2). For six reference points,

Table 2: Chi-square (X^2) test with Friedman method for finding difference among six infilling methods. SD represents standard deviation. P value means significance probability.

(a) 4-NN

	N	Mean	SD	Min.	Max. value	Mean rank	X^2	P value
Ep	19	−1.29	4.46	−15.50	5.76	2.53	55.602	0.0000
Qu	19	−1.42	5.03	−15.06	7.51	2.74		
Tw	19	−1.64	5.37	−14.90	8.13	2.58		
Tc	19	−1.18	5.06	−14.87	8.29	4.47		
Co	19	−1.40	4.69	−15.42	6.08	2.68		
Reg	19	2.76	5.42	−13.47	14.38	6.00		

(b) 6-NN

	N	Mean	SD	Min.	Max. value	Mean rank	X^2	P value
Ep	19	−2.61	4.72	−16.68	1.58	1.53	66.519	0.0000
Qu	19	−2.27	4.71	−16.20	2.82	3.16		
Tw	19	−2.18	4.84	−15.89	4.06	3.79		
Tc	19	−2.15	4.63	−16.20	2.84	4.21		
Co	19	−2.54	4.69	−16.59	1.50	2.32		
Reg	19	0.06	4.97	−15.48	7.33	6.00		

(c) 8-NN

	N	Mean	SD	Min.	Max. value	Mean rank	X^2	P value
Ep	19	−3.40	5.04	−17.33	1.94	1.32	75.812	0.0000
Qu	19	−3.10	4.75	−16.90	0.45	3.21		
Tw	19	−2.74	4.79	−16.59	1.29	4.58		
Tc	19	−2.93	4.77	−16.96	1.51	3.68		
Co	19	−3.28	4.93	−17.25	1.85	2.21		
Reg	19	−1.24	5.35	−16.49	8.08	6.00		

(Ep: Epanechnikov, Qu: Quartic, Tw: Triweight, Tc: Tricube, Co: Cosine, and Reg: regression).

the $k$nn-regression, Tricube, Triweight, Quartic, Cosine, and Epanechnikov were ranked as shown in Table 2. In another example, eight reference points used $k$nn-regression, Triweight, Quartic, Cosine, and Epanechnikov average rank (Table 2). As shown in Table 2, the $k$nn-regression has the largest average rank and Epanechnikov has the smallest rank average for all of the reference point cases. This result proves the dissimilarity of these methods.

To determine which methods are dissimilar to the others, this study performed the Wilcoxon signed rank test [38]. The basic feature of the Wilcoxon signed rank test is that data samples that come from the same population are paired and it is detailed in the following:

$$W = \left| \sum_{i=1}^{N} \left[\operatorname{sign}\left(y_{2,i} - y_{1,i}\right) R_i \right] \right|, \qquad (10)$$

where N is the sample size, $y_{2,i}$ is ith value of the second data point, $y_{1,i}$ is ith value of the first data point, and R_i is the rank of $|y_{2,i} - y_{1,i}|$. If the W value is less than 5%, it means there is different mechanism used on the sample data or method. Table 3 shows that the W value for $k$nn-regression is less than 5% for all cases. Accordingly, this signifies that $k$nn-regression is completely dissimilar to the other methods. Although the five different kernel methods for data interpolation exhibit similarity or dissimilarity to each other depending on the number of reference points, all of the kernel methods can be distinguished from $k$nn-regression using the Wilcoxon signed rank test.

4. Results

Since Epanechnikov has the smallest average rank, which signifies a small difference between the observation value and the interpolated value for all reference points in Table 2, interpolation data obtained from the Epanechnikov method has the best result among the studied methods. Figure 2 shows that filling in data from $k$nn-regression has a large difference at both four and six reference points. Interpolation data from the kernel methods are close to zero for both the average and median values at four reference points, meaning that the interpolation data are similar to the observation data. On the other hand, more than 75% of the interpolation data from $k$nn-regression exhibits a difference than zero. When the interpolation data are evaluated at six reference points in Figure 2, the median value from the $k$nn-regression is shown to be far away from zero. At eight reference points, $k$nn-regression is close to zero for both average and median values; however, it is difficult to conclude that this is an ideal method

Table 3: Chi-square (X^2) test with Wilcoxon signed rank method between regression and five different kernel methods.

(a) 4-NN

		Ep	Qu	Tw	Tc	Co
Reg	X^2	−3.823	−3.823	−3.823	−3.823	−3.823
	P value	0.0001	0.0001	0.0001	0.0001	0.0001

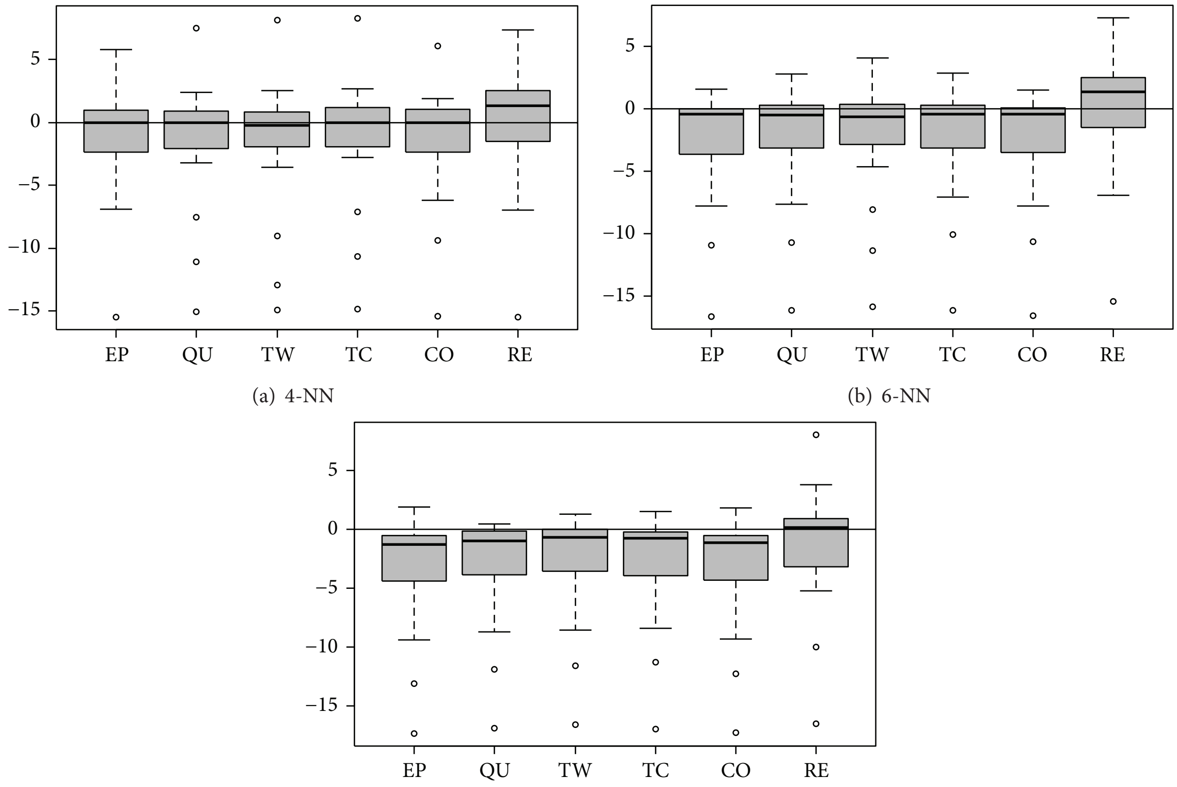
(b) 6-NN

		Ep	Qu	Tw	Tc	Co
Reg	X^2	−3.823	−3.823	−3.823	−3.823	−3.823
	P value	0.0001	0.0001	0.0001	0.0001	0.0001

(c) 8-NN

		Ep	Qu	Tw	Tc	Co
Reg	X^2	−3.823	−3.823	−3.823	−3.823	−3.823
	P value	0.0001	0.0001	0.0001	0.0001	0.0001

(Ep: Epanechnikov, Qu: Quartic, Tw: Triweight, Tc: Tricube, Co: Cosine, and Reg: regression).



(a) 4-NN

(b) 6-NN

(c) 8-NN

Figure 2: Box plots for difference between actual precipitation and interpolated precipitation. y-axis represents mm per day.

because outlying maximum values will affect the average and median value.

This study on precipitation data interpolation also evaluated the simulation of the interpolated data using the SWAT hydrologic model. In SWAT hydrologic modeling, the surface runoff is estimated by considering excess precipitation with abstractions and infiltration factor through Soil Conservation Service Curve Number (SCS-CN) method. Green-Ampt (GA) infiltration method is another method to calculate the surface runoff in SWAT. A study shows that both methods give reasonable results, and there is no significant advantage observed in using one over the other. However, the GA method appears to have more limitations in modeling seasonal variability than the SCS-CN method does. Hence, the SCS-CN method is used for infiltration factor in this study. An SCS curve number based simulation needs

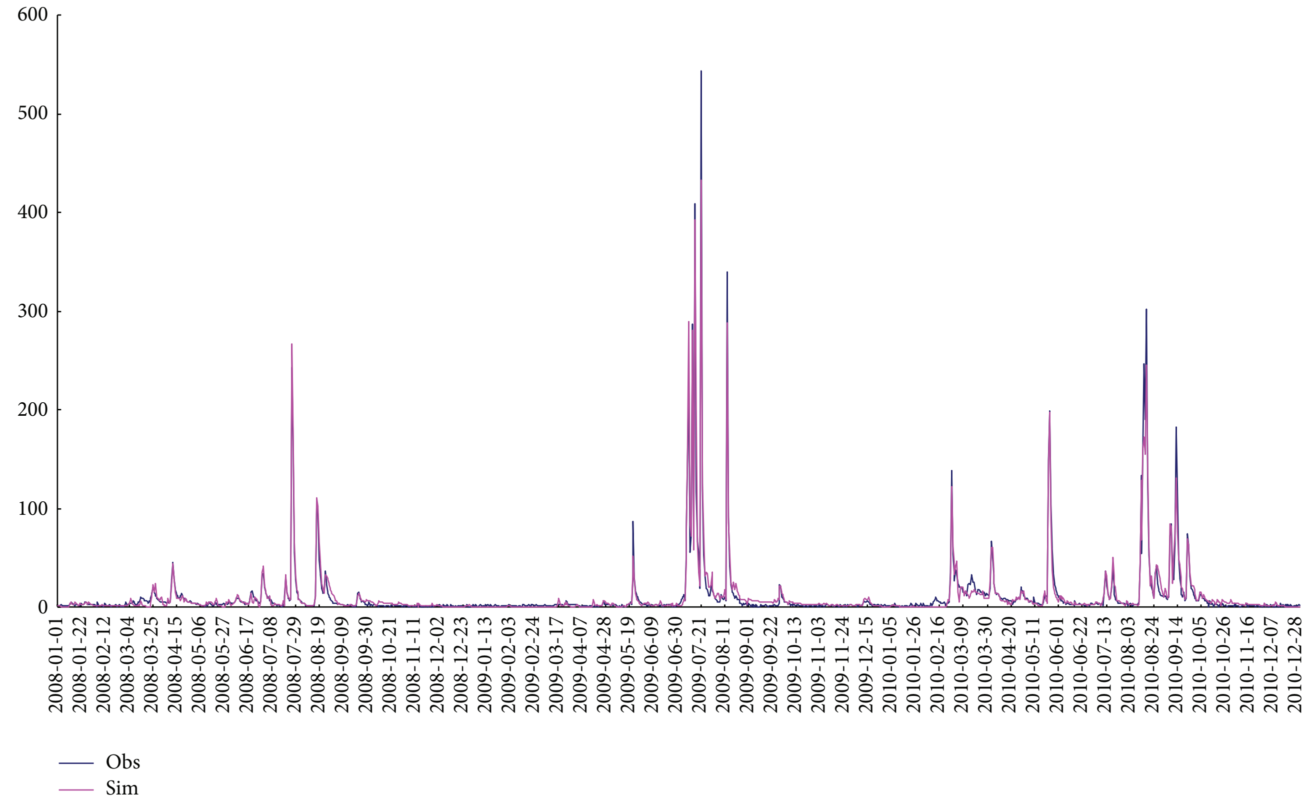


Figure 3: Calibrated model result using original precipitation input. x-axis represents time in days and y-axis represents flow in cubic meters per second.

Table 4: Details of SWAT parameters which are related to runoff mechanism for Imha watershed.

Parameter	Description	Selected value
ESCO	Soil evaporation compensation factor	0.9500
EPCO	Plant water uptake compensation factor	1.0000
EVLAI	Leaf area index at which no evaporation occurs from water surface [m^2/m^2]	3.0000
FFCB	Initial soil water storage expressed as a fraction of field capacity water content	0.0000
IEVENT	Rainfall/runoff code: 0 = daily rainfall/CN	0.0000
ICRK	Crack flow code: 1 = model crack flow in soil	0.0000
SURLAG	Surface runoff lag time [days]	4.0000
ADJ_PKR	Peak rate adjustment factor for sediment routing in the subbasin (tributary channels)	0.0000
PRF	Peak rate adjustment factor for sediment routing in the main channel	1.0000
SPCON	Linear parameter for calculating the maximum amount of sediment that can be reentrained during channel sediment	0.0001
SPEXP	Exponent parameter for calculating sediment reentrained in channel sediment routing	1.0000

time step updated information as soil water content changes. Excess rainfall equation in SCS-CN method was generated based on historical relationship between the curve number and the hydrologic mechanism for over 20 years. Throughout the surface runoff calculation, infiltration should be updated over time according to the soil type. Other abstractions such as evapotranspiration and soil and snow evaporation are calculated by Penman-Monteith method and meteorological statistics. Finally, the kinematic storage model is used to compute groundwater storage and seepage. Flow resulting in SWAT modeling is routed HRUs to watershed outlet. Figure 3 shows the calibration of the model simulation as the initial step and the specific parameters are described in Table 4. After the calibration of the SWAT model, the six different interpolated precipitation datasets, with three different reference ranges for each (a total of twenty-four

Table 5: Details of simulation results with six different precipitation infilling methods in Imha watershed.

	4-NN			6-NN			8-NN		
	E_{NS}	R^2	RMSE	E_{NS}	R^2	RMSE	E_{NS}	R^2	RMSE
Ep	0.80	0.83	15.32	0.91	0.92	10.60	0.78	0.82	16.16
Qu	0.73	0.78	17.83	0.91	0.93	10.48	0.88	0.92	11.80
Tw	0.91	0.91	10.56	0.93	0.94	9.25	0.95	0.95	8.10
Tc	0.95	0.95	7.72	0.93	0.94	9.03	0.95	0.95	7.64
Co	0.93	0.94	8.83	0.95	0.95	7.72	0.91	0.93	10.44
Reg	0.69	0.80	19.14	0.21	0.65	30.71	0.71	0.73	21.48

interpolated precipitations data points), were used to assess the performance of interpolated precipitation data for hydrologic model simulation. Streamflow simulations were done for three years from 2008 to 2010. To evaluate the model performance considering the use of different interpolated precipitation datasets, this study used E_{NS} (Nash-Sutcliffe coefficient), R-square (coefficient of determination), and RMSE (root mean square error). Table 5 and Figure 4 show that the simulation results from $k$nn-regression exhibit low SWAT simulation performance for streamflow estimations, with 0.54 E_{NS}, 0.74 R-square, and 23.78 m^3/s RMSE as an average. All of the kernel functions, on the other hand, exhibit good performance for hydrologic simulations with interpolated precipitation data (Table 5 and Figure 4), the average of E_{NS}, R-square, and RMSE (1) for Epanechnikov is 0.83, 0.86, and 14.03 m^3/s; (2) for Quartic is 0.84, 0.88, and 13.03 m^3/s; (3) for Triweight is 0.93, 0.93, and 9.30 m^3/s; (4) for Tricube is 0.94, 0.95, and 8.13 m^3/s; and (5) for Cosine is 0.93, 0.94, and 9.00 m^3/s, respectively.

5. Conclusions

Five different kernel functions were applied to the Imha watershed to evaluate the performance of each weighted method for estimating missing precipitation data and the use of interpolated data for hydrologic simulations was assessed. The following conclusions can be drawn from this research.

(1) To estimate missing precipitation data points, exploratory procedures should consider the spatiotemporal variations of precipitation. Due to difficulty on accounting for these variations, statistical methods for estimating missing precipitation data are commonly used.

(2) Although ANNs are an advanced approach for estimating missing data, mechanisms are unclear because the neuron system is ultimately a black-box model. Thus, regression methods are widely used for estimating missing data, even though there are limitations in that regression methods cannot follow normal distribution when the sample is small.

(3) When using kernel functions as a weighted method, estimated missing data would satisfy normal distribution which is more statistically sound. Also, kernel methods can overcome weakness in $k$nn-regression if the data have outliers and/or a nonlinear trend around the missing data points in terms of mean value.

(4) This study assessed the five kernel functions, Epanechnikov, Quartic, Triweight, Tricube, and Cosine, as a weight for predicting missing values. In comparison with the $k$nn-regression method, this study demonstrates that the kernel approaches provide higher quality interpolated precipitation data than the $k$nn-regression approach. In addition, the kernel function results better conform to statistical standards.

(5) Furthermore, higher quality of interpolated precipitation data results in better performance for hydrologic simulations, as exemplified in this study. All of the statistical analyses of the streamflow simulations showed that the simulations using the interpolated precipitation data from the kernel functions provide better results than using $k$nn-regression.

(6) Use of kernel distribution is a more effective method than regression when the precipitation data have an upward or downward trend. However, if the precipitation data have a nonlinear trend, it is difficult to effectively reconstruct the missing values. For further research, a time series analysis or a random walk model using a stochastic process are possible methods by which to estimate missing data where there is a nonlinear trend.

Appendices

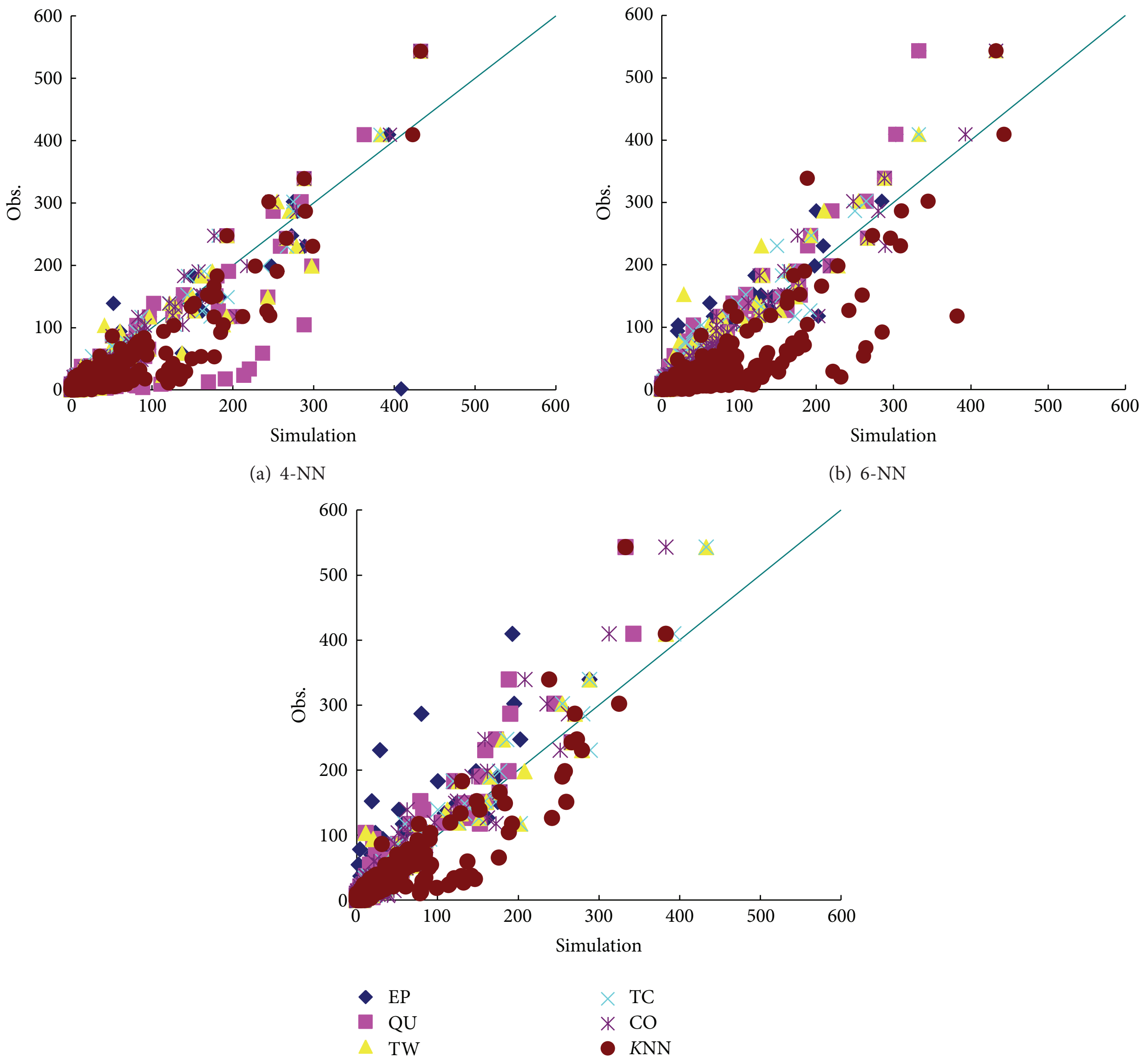
A. Kenel Functions

Kernel density estimation is an unsupervised learning procedure, which historically precedes kernel regression. It also leads naturally to a simple family of procedures for nonparametric classification.

A.1. Kernel Density Estimation. Suppose we have a random sample $x_1, x_2, \ldots, x_N$ draw from a probability density $f_x(x)$ and we wish to estimate f_x at a point x_0. For simplicity we assume for now that $x \in R$ (*real value*). Arguing as before, a natural local estimate has the form

$$\hat{f}_x(x_0) = \frac{\# x_i \in N(x_0)}{N\lambda}, \quad \text{(A.1)}$$

where $\# x_i$ means number of x_i which converges to $N(x_0)$ and $N(x_0)$ is a small metric neighborhood around x_0 of width λ.



(a) 4-NN

(b) 6-NN

(c) 8-NN

Figure 4: Scatter plots for SWAT simulation (EP, QU, TW, TC, CO, and *K*NN represents Epanechnikov, Quartic, Triweight, Tricube, Cosine, and *K*NN-regression, resp.).

This estimate is bumpy, and the smooth Parzen estimate is preferred,

$$\hat{f}_x(x_0) = \frac{1}{N\lambda}\sum_{i=1}^{N} K_\lambda(x_0, x_i), \qquad \text{(A.2)}$$

because it counts observations close to x_0 with weights that decrease with distance from x_0. In this case a popular choice for K_λ is the Gaussian kernel $K_\lambda(x_0, x_i) = \phi(|x - x_0|/\lambda)$. Letting ϕ_λ denote the Gaussian density with mean zero and standard-deviation λ, then (A.2) has the form

$$\hat{f}_x(x_0) = \frac{1}{N}\sum_{i=1}^{N} \phi_\lambda(x - x_i) = \left(\hat{F} * \phi_\lambda\right)(x), \qquad \text{(A.3)}$$

the convolution of the sample empirical distribution $\hat{F}$ with ϕ_λ. The distribution $\hat{F}(x)$ puts mass $1/N$ at each of the observed x_i and is jumpy; in $\hat{f}_x(x)$ we have smoothed $\hat{F}$ by adding independent Gaussian noise to each observation x_i.

The Parzen density estimate is the equivalent of the local average, and improvements have been proposed along the lines of local regression (on the log scale for densities). We will not pursue these here. In R^p the natural generalization of the Gaussian density estimate amounts to using the Gaussian product kernel in (A.3),

$$\hat{f}_x(x_0) = \frac{1}{N(2\lambda^2\pi)^{p/2}}\sum_{i=1}^{N} e^{-(1/2)(\|x_i - x_0\|/\lambda)^2}. \qquad \text{(A.4)}$$

Table 6: Weighted values depending on day distance with each KNN.

	4-NN		6-NN			8-NN			
	1st	2nd	1st	2nd	3rd	1st	2nd	3rd	4th
Ep	0.667	0.417	0.703	0.563	0.328	0.720	0.630	0.480	0.270
Qu	0.741	0.289	0.824	0.527	0.179	0.864	0.662	0.384	0.122
Tw	0.768	0.188	0.901	0.461	0.092	0.968	0.648	0.287	0.051
Tc	0.772	0.301	0.824	0.579	0.167	0.844	0.709	0.416	0.100
Co	0.680	0.393	0.726	0.555	0.301	0.747	0.635	0.462	0.243

A.2. Kernel Density Classification. One can use nonparametric density estimates for classification in a straight-forward fashion using Bayes' theorem. Suppose for a J class problem we fit nonparametric density estimates $\hat{f}_j(X)$, $j = 1, \ldots, J$ separately in each of the classes, and we also have estimates of the class priors $\hat{\pi}_j$ (usually the sample proportions). Then

$$\widehat{\Pr}\left(G = j \mid X = x_0\right) = \frac{\hat{\pi}_j \hat{f}_j\left(x_0\right)}{\sum_{k=1}^{J} \hat{\pi}_k \hat{f}_k\left(x_0\right)}. \qquad (\text{A.5})$$

In this region the data are sparse for both classes, and since the Gaussian kernel density estimates use matric kernels, the density estimates are low and of poor quality (high variance) in these regions. The local logistic regression method uses the tricube kernel with k-NN bandwidth; this effectively widens the kernel in this region and makes use of the local linear assumption to smooth out the estimate (on the logit scale).

If classification is the ultimate goal, then learning the separate class densities well may be unnecessary and can in fact be misleading. In learning the separate densities form data, one might decide to settle for a rougher, high-variance fit to capture these features, which are irrelevant for the purposes of estimating the posterior probabilities. In fact, if classification is the ultimate goal, then we need only to estimate the posterior well near the decision boundary (for two classes, this is the set $\{x \mid \Pr(G = 1 \mid X = x) = 1/2\}$).

B. Procedures of Missing Precipitation

This step shows example calculation for kernel functions for weighted mean. It is an example question about the weight of each situation. If the kernel functions are all symmetric, same values are used for weight based on day distance. Following Table 6 1st, 2nd, 3rd, and 4th day distance and weighted values are shown. For example, if we want to estimate missing precipitation for 2010-02-12 (actual value is 6), see following procedures (3 steps) with 4-NN Epanechnikov kernel (Table 7).

Step 1. Select the date for target interpolation data.

Step 2. Decide Kth nearest days precipitation and each kernel weight.

Step 3. Calculate the weight average to estimate missing.

Table 7: Example calculation to interpolate for missing precipitation.

Step 1		Step 2		Step 3
Date	Prec.	Weight	Prec.·Weight	Estimation
2010-02-10	15	0.417	6.255	5.949
2010-02-11	17.2	0.667	11.472	
2010-02-12	6	—	—	
2010-02-13	9.1	0.667	6.070	
2010-02-14	0	0.417	0	

Table 8: Calculating missing precipitation for 2010-02-12 (actual value is 6) with six different methods.

Method	4NN	6NN	8NN
Ep	5.949	5.172	4.862
Qu	5.956	5.302	4.936
Tw	5.755	5.294	4.952
Tc	6.205	5.407	4.963
Co	5.945	5.197	4.876
Reg	10.325	8.967	8.813

(Ep: Epanechnikov, Qu: Quartic, Tw: Triweight, Tc: Tricube, Co: Cosine, and Reg: regression).

The rest of the kernel methods for estimating missing precipitation are described in Table 8.

C. Sample Calculations with Real Value

This section shows how to calculate missing precipitation with kernel mean weighed function by using certain number. This sample selected daily data from 2008 to 2010 with 0.02 possibilities to bivariate by random. After selected data, setting data location is operated. Zhang et al. [39] addressed that kernel based nonparametric multiple imputation has better performance than general linear regression when the sample data is small or limited.

Table 9 shows procedure of kernel weight in each function. We used data Feb. 10, 2012 from Feb. 14, 2014 to estimate Feb. 12, 2012 missing data. Epanechnikov kernel showed that longest data has highest estimation as 0.417; however, Triweight kernel showed that longest data has lowest estimation as 0.188. Highest weight in nearest value is Tricube kernel and lowest weight is Epanechnikov kernel. Generally,

Table 9: Sample calculation with certain number.

(a) Epanechnikov

Date	2.10.	2.11.	2.12.	2.13.	2.14.
Prec.	15.0	17.2	6.0	9.1	0.0
Ep. weight	0.417	0.667	—	0.667	0.417
Prec.·weight	6.26	11.47	—	6.07	0.00
Estimation			5.95		

(b) Quartic

Date	2.10.	2.11.	2.12.	2.13.	2.14.
Prec.	15.0	17.2	6.0	9.1	0.0
Qu. weight	0.289	0.741	—	0.741	0.289
Prec.·weight	4.34	12.75	—	6.74	0.00
Estimation			5.96		

(c) Triweight

Date	2.10.	2.11.	2.12.	2.13.	2.14.
Prec.	15.0	17.2	6.0	9.1	0.0
Tw. weight	0.188	0.768	—	0.768	0.188
Prec.·weight	2.82	13.21	—	6.99	0.00
Estimation			5.75		

(d) Tricube

Date	2.10.	2.11.	2.12.	2.13.	2.14.
Prec.	15.0	17.2	6.0	9.1	0.0
Tc. weight	0.301	0.772	—	0.772	0.301
Prec.·weight	4.52	13.28	—	7.03	0.00
Estimation			6.20		

(e) Cosine

Date	2.10.	2.11.	2.12.	2.13.	2.14.
Prec.	15.0	17.2	6.0	9.1	0.0
Co. weight	0.393	0.680	—	0.680	0.393
Prec.·weight	5.90	11.70	—	6.19	0.00
Estimation			5.94		

Tricube, that is, high weight, shows the overestimation for missing precipitation.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] K. Kang and V. Merwade, "Development and application of a storage-release based distributed hydrologic model using GIS," *Journal of Hydrology*, vol. 403, no. 1-2, pp. 1–13, 2011.

[2] A. J. Abebe, D. P. Solomatine, and R. G. W. Venneker, "Application of adaptive fuzzy rule-based models for reconstruction of missing precipitation events," *Hydrological Sciences Journal*, vol. 45, no. 3, pp. 425–436, 2000.

[3] P. Ramos-Calzado, J. Gómez-Camacho, F. Pérez-Bernal, and M. F. Pita-López, "A novel approach to precipitation series completion in climatological datasets: application to Andalusia," *International Journal of Climatology*, vol. 28, no. 11, pp. 1525–1534, 2008.

[4] T. Schneider, "Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values," *Journal of Climate*, vol. 14, no. 5, pp. 853–871, 2001.

[5] S. K. Regonda, D.-J. Seo, B. Lawrence, J. D. Brown, and J. Demargne, "Short-term ensemble stream forecasting using operationally-produced single-valued streamflow forecasts—a Hydrologic Model Output Statistics (HMOS) approach," *Journal of Hydrology*, vol. 497, pp. 80–96, 2013.

[6] P. Coulibaly and N. D. Evora, "Comparison of neural network methods for infilling missing daily weather records," *Journal of Hydrology*, vol. 341, no. 1-2, pp. 27–41, 2007.

[7] H. A. El Sharif and R. S. V. Teegavarapu, "Evaluation of spatial interpolation methods for missing precipitation data: preservation of spatial statistics," in *Proceedings of the World Environmental and Water Resources Congress*, pp. 3822–3832, May 2012.

[8] A. Bárdossy and G. Pegram, "Infilling missing precipitation records—a comparison of a new copula-based method with other techniques," *Journal of Hydrology*, vol. 519, pp. 1162–1170, 2014.

[9] K. Schamm, M. Ziese, A. Becker et al., "Global gridded precipitation over land: a description of the new GPCC first guess daily product," *Earth System Science Data*, vol. 6, no. 1, pp. 49–60, 2014.

[10] R. S. V. Teegavarapu, "Statistical corrections of spatially interpolated missing precipitation data estimates," *Hydrological Processes*, vol. 28, no. 11, pp. 3789–3808, 2014.

[11] Y. Da and G. Xiurun, "An improved PSO-based ANN with simulated annealing technique," *Neurocomputing*, vol. 63, pp. 527–533, 2005.

[12] A. di Piazza, F. L. Conti, L. V. Noto, F. Viola, and G. La Loggia, "Comparative analysis of different techniques for spatial interpolation of rainfall data to create a serially complete monthly time series of precipitation for Sicily, Italy," *International Journal of Applied Earth Observation and Geoinformation*, vol. 13, no. 3, pp. 396–408, 2011.

[13] E. Pisoni, F. Pastor, and M. Volta, "Artificial Neural Networks to reconstruct incomplete satellite data: application to the Mediterranean Sea Surface Temperature," *Nonlinear Processes in Geophysics*, vol. 15, no. 1, pp. 61–70, 2008.

[14] V. Sharma, S. Rai, and A. Dev, "A comprehensive study of artificial neural networks," *International Journal of Advanced Research in Computer Science and Sofrware Engineering*, vol. 2, no. 10, pp. 278–284, 2012.

[15] ASCE Task Committee, "Artificial neural networks in hydrology. II: hydrologic applications," *Journal of Hydrological Engineering*, vol. 5, no. 2, pp. 124–137, 2000.

[16] ASCE Task Committee, "Artificial neural networks in hydrology. I: preliminary concepts," *Journal of Hydrologic Engineering*, vol. 5, no. 2, pp. 115–123, 2000.

[17] D. E. Rumelhart, B. Widrow, and M. A. Lehr, "The basic ideas in neural networks," *Communications of the ACM*, vol. 37, no. 3, pp. 87–92, 1994.

[18] J. Amorocho and W. E. Hart, "A critique of current methods in hydrologic systems investigation," *Transactions of the American Geophysical Union*, vol. 45, no. 2, pp. 307–321, 1964.

[19] A. W. Minns and M. J. Hall, "Artificial neural networks as rainfall-runoff models," *Hydrological Sciences Journal*, vol. 41, no. 3, pp. 399–417, 1996.

[20] K. Kang and V. Merwade, "The effect of spatially uniform and non-uniform precipitation bias correction methods on improving NEXRAD rainfall accuracy for distributed hydrologic modeling," *Hydrology Research*, vol. 45, no. 1, pp. 23–42, 2014.

[21] C. Daly, W. P. Gibson, G. H. Taylor, G. L. Johnson, and P. Pasteris, "A knowledge-based approach to the statistical mapping of climate," *Climate Research*, vol. 22, no. 2, pp. 99–113, 2002.

[22] J. D. Creutin, H. Andrieu, and D. Faure, "Use of a weather radar for the hydrology of a mountainous area. Part II: radar measurement validation," *Journal of Hydrology*, vol. 193, no. 1–4, pp. 26–44, 1997.

[23] Y. L. Xia, P. Fabian, A. Stohl, and M. Winterhalter, "Forest climatology: estimation of missing values for Bavaria, Germany," *Agricultural and Forest Meteorology*, vol. 96, no. 1–3, pp. 131–144, 1999.

[24] C. J. Willmott, S. M. Robeson, and J. J. Feddema, "Estimating continental and terrestrial precipitation averages from raingauge networks," *International Journal of Climatology*, vol. 14, no. 4, pp. 403–414, 1994.

[25] R. S. V. Teegavarapu and V. Chandramouli, "Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records," *Journal of Hydrology*, vol. 312, no. 1–4, pp. 191–206, 2005.

[26] J. A. Smith, "Precipitation," in *Handbook of Hydrology*, D. R. Maidment, Ed., vol. 3, chapter 3, McGraw Hill, New York, NY, USA, 1993.

[27] J. R. Simanton and H. B. Osborn, "Reciprocal-distance estimate of point rainfall," *Journal of Hydraulic Engineering Division*, vol. 106, no. 7, pp. 1242–1246, 1980.

[28] J. D.-J. Salas, "Analysis and modeling of hydrological time series," in *Handbook of Hydrology*, D. R. Maidment, Ed., vol. 19, chapter 19, pp. 19.1–19.72, McGraw-Hill, New York, NY, USA, 1993.

[29] S. J. Jeffrey, J. O. Carter, K. B. Moodie, and A. R. Beswick, "Using spatial interpolation to construct a comprehensive archive of Australian climate data," *Environmental Modelling and Software*, vol. 16, no. 4, pp. 309–330, 2001.

[30] M. Franklin, V. R. Kotamarthi, M. L. Stein, and D. R. Cook, "Generating data ensembles over a model grid from sparse climate point measurements," *Journal of Physics: Conference Series*, vol. 125, Article ID 012019, 2008.

[31] K. C. Young, "A three-way model for interpolating for monthly precipitation values," *Monthly Weather Review*, vol. 120, no. 11, pp. 2561–2569, 1992.

[32] F. Filippini, G. Galliani, and L. Pomi, "The estimation of missing meteorological data in a network of automatic stations," *Transactions on Ecology and the Environment*, vol. 4, pp. 283–291, 1994.

[33] W. P. Lowry, *Compendium of Lecture Notes in Climatology for Class IV Meteorological Personnel*, Secretariat of the World Meteorological Organization, Geneva, Switzerland, 1972.

[34] M. C. Acock and Y. A. Pachepsky, "Estimating missing weather data for agricultural simulations using group method of data handling," *Journal of Applied Meteorology*, vol. 39, no. 7, pp. 1176–1184, 2000.

[35] S. L. Neitsch, J. G. Arnold, J. R. Kiniry, J. R. Williams, and K. W. King, *Soil and Water Assessment Tool—Theoretical Documentation (Version 2005)*, Texas Water Resource Institute, College Station, Tex, USA, 2005.

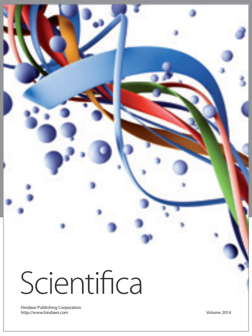
[36] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality: complete samples," *Biometrika*, vol. 52, no. 3-4, pp. 591–611, 1965.

[37] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American Statistical Association*, vol. 32, no. 200, pp. 675–701, 1937.

[38] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

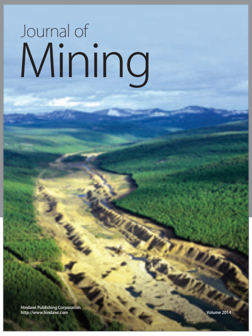
[39] S. Zhang, Z. Jin, X. Zhu, and J. Zhang, "Missing data analysis: a kernel-based multi-imputation approach," in *Transactions on Computational Science III*, vol. 5300 of *Lecture Notes in Computer Science*, pp. 122–142, Springer, Berlin, Germany, 2009.

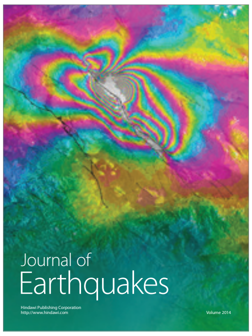
Journal of
Geochemistry


Scientifica

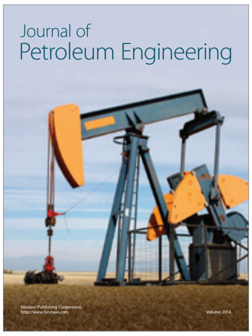
International Journal of
Ecology

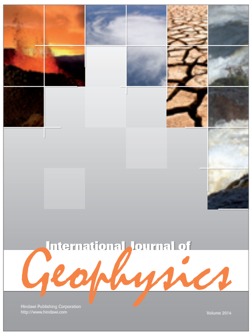
The Scientific
World Journal

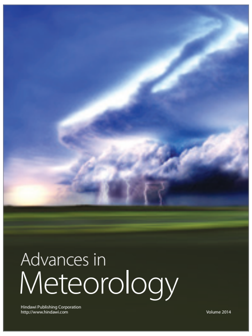

Journal of
Mining

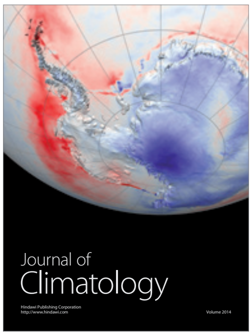

Journal of
Earthquakes

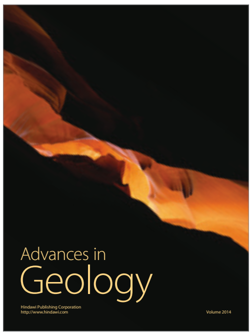
Paleontology Journal

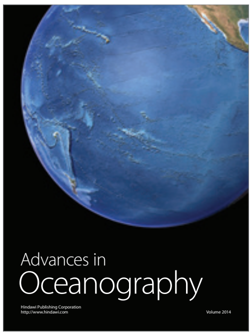

Journal of
Petroleum Engineering

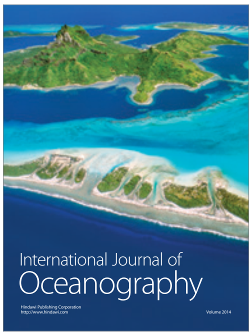

International Journal of
Geophysics


Advances in
Meteorology


Journal of
Climatology

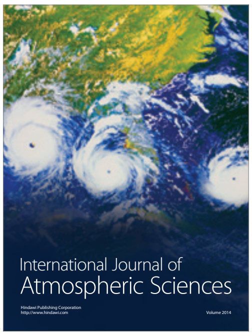

Advances in
Geology

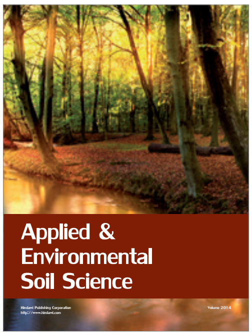

Advances in
Oceanography


International Journal of
Oceanography

International Journal of
Mineralogy

Journal of
Geological Research


International Journal of
Atmospheric Sciences


Applied &
Environmental
Soil Science

Journal of
Computational
Environmental Sciences