Journal of Hydrology 394 (2010) 305–314

Contents lists available at ScienceDirect

Journal of Hydrology

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

Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation

Jung-Woo Kim *, Yakov A. Pachepsky

USDA-ARS, Environmental Microbial & Food Safety Laboratory, 10300 Baltimore Ave., Beltsville, MD 20705, USA

ARTICLE INFO

Article history:
Received 2 April 2010
Received in revised form 16 August 2010
Accepted 5 September 2010

This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Fi-John Chang, Associate Editor

Keywords:
Artificial neural network
Precipitation
Reconstruction missing data
Regression tree
Streamflow

SUMMARY

Environmental models typically require a complete time series of meteorological inputs, thus reconstructing missing data is a key issue in the functionality of such physical models. The objective of this work was to develop a new technique to reconstruct missing daily precipitation data in the central part of Chesapeake Bay Watershed using a two-step reconstruction method (RT + ANN) that employed artificial neural networks (ANN) with inputs only from stations that were found to be influential in bootstrap applications of regression trees (RT). The predictive performance of RT + ANN was also compared with those of stand-alone RT and ANN methods. In addition to statistical comparisons of the reconstructed precipitation time series, these resulting data in the Soil and Water Assessment Tool (SWAT) watershed model were used to perform an error propagation analysis in streamflow simulations. The RT provided a transparent visual representation of the similarity between the stations in their daily precipitation time series. Seven years of data from 39 weather stations showed that both RT and ANN provided the reconstruction accuracy comparable to (or better than) published earlier results of precipitation reconstruction. The RT + ANN method significantly improved accuracy and was more robust when compared to RT or ANN methods. This method also provided more accurate and robust SWAT streamflow predictions with reconstructed precipitation.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

Environmental models typically require a complete time series of meteorological inputs, thus reconstructing missing data is a key issue in the functionality of such physical models. Weather stations in the watershed sometimes malfunction during the monitoring period. Missing data has been a problem and therefore have to be estimated for models to function. Estimation of missing data, also known as infilling (Coulibaly and Evora, 2007), reconstruction (Abebe et al., 2000), meteorological data series completion (Ramos-Calzado et al., 2008), or imputation of meteorological data (Schneider, 2001), has been a topic of interest in many hydrologic and other environmental modeling studies.

The estimation of missing meteorological data is typically based on data for the same location on days when data is available or on data from neighboring stations for the same day. When data from the same station is used, the reconstruction methods include a simple interpolation between available data or their derivatives (Lowry, 1972), using a mean value of the data series (Linacre, 1992), or using data from several days before and several days after the date of missing data in a non-linear regression to fill the 1-day data gap (Acock and Pachepsky, 2000). Observations from neighboring stations are used in missing data reconstruction much more often. It has been proposed to use geometrical distances to the stations and to employ the data from the closest station (Xia et al., 1999), the arithmetic averaging of data from several neighboring stations (Willmott et al., 1994), and the inverse distance weighting method (Teegavarapua and Chandramouli, 2005). Statistical interpolation methods, such as kriging (Jeffrey et al., 2001) or spatio-temporal interpolation (Franklin et al., 2008) are typically applied in the estimation. Ramos-Calzado et al. (2008) infilled missing monthly precipitation data by using linear regressions and estimating the missing value as a median of the distribution of the regression outputs. In addition to using statistical similarity between stations to define weights for spatial interpolation, Young (1992) and Filippini et al. (1994) proposed to apply spatial interpolation using correlations between time series of stations to assign weights.

The accuracy of the reconstruction differs among weather variables and has been the lowest for precipitation. This is due to the highly stochastic nature of precipitation being relatively compared to other climatic variables such as temperature, which show a

* Corresponding author. Present address: Radioactive Waste Technology Development Division, Korea Atomic Energy Research Institute, 1045 Daedeok-daero, Yuseong-gu, Daejeon 305–353, Republic of Korea. Tel.: +82 42 868 2547; fax: +82 42 868 8850.

E-mail address: jw_kim@kaeri.re.kr (J.-W. Kim).

0022-1694/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhydrol.2010.09.005

lower degree of spatial and temporal variability. Application of the interpolation within time series to highly variable meteorological data, such as precipitation, was limited (Ramos-Calzado et al., 2008). Reconstruction of the missing climatological data using neighboring stations has also been the least successful for precipitation (Xia et al., 1999). This happens probably because the complex effects of topography and climatology on spatial distribution of precipitation are not well known and difficult to incorporate into the precipitation imputation algorithms (Daly et al., 1994; Thornton et al., 1997; Xia et al., 1999). Artificial neural networks (ANNs) as the non-linear statistical data modeling tools capable to model complex relationships between inputs and outputs have recently been used to reconstruct missing weather data (Kuligowski and Barros, 1998; Michaelides et al., 2007). Coulibaly and Evora (2007) compared six different types of artificial neural networks by their applicability for infilling missing daily total precipitation values.

The trend in missing data reconstruction has been to derive the missing data estimates from a time series that exhibit some similarity with the time series in question. The preliminary grouping with cluster analysis (Garcia et al., 2006) and the nearest neighbors approach (Siripitayananon et al., 2003) have been employed to select the best predictors of missing rainfall and wind speed data, respectively. Regression trees have been shown to be an efficient technique for establishing the transparent and explicable set of the best predictors, and as such have been widely employed in atmospheric studies (e.g., Burrows et al., 1995). Faucher et al. (1999) successfully used regression trees to select predictors to estimate marine winds.

The interest in this study in missing data reconstruction stemmed from the need of applying Soil and Water Assessment Tool (SWAT, Santhi et al., 2006; Krysanova and Arnold, 2008) to evaluate the effect of landuse on bacterial water quality in a small agricultural watershed in Southern Pennsylvania (Kim et al., 2010). SWAT, the watershed-scale physically-based model, has been developed to simulate the fate and transport of pollutants coming from nonpoint sources (Neitsch et al., 2005).

The objective of this work was to develop a precipitation data reconstruction technique that would combine the advantages of regression tree and artificial neural network. The hypothesis was that reconstruction of missing precipitation data will be more efficient if the influential input variables are determined by using the regression tree technique and used in artificial neural networks. The reconstruction methods were compared using bootstraps to obtain statistical distributions of the performance statistics – the correlation coefficient and the root mean square error. The reconstruction methods were also evaluated by comparing results of SWAT streamflow simulations with the reconstructed precipitation data.

2. Materials and methods

2.1. Study area and precipitation data

The study area is located in the central part of Chesapeake Bay Watershed between 38°53′–40°41′N and 76°43′–79°08′W. It includes the Appalachian Mountains in the western part and plain urbanized area in the eastern part, and crossed by the Potomac and Susquehanna Rivers. Daily precipitation data were downloaded from the US National Climatic Data Center's database (NCDC, 2008) for 39 weather stations within the study area for 7 years from January 1, 2001 to December 31, 2007. For the preliminary, percentages of valid daily precipitation data (%) and average annual precipitation (mm) for each station during the 7 years were estimated. In addition, the average probability of non-zero daily precipitation (%) was calculated by examining the number of days of precipitation per year and taking the average of these annual proportions over the study period.

All weather stations accumulated daily values over a 24-h period, but the ending integration times differed among stations (Table 1). In order to ensure data compatibility, the daily precipitation data of each weather station were corrected based on the ending integration time (t_d, unit: hour):

$$P'_d = P_d \cdot \left(\frac{t_d}{24}\right) + P_{d+1} \cdot \left(1 - \frac{t_{d+1}}{24}\right) \quad (1)$$

where P'_d was the corrected daily precipitation on day d, and P_d and P_{d+1} were the reported daily precipitations on days d and $d+1$, respectively. Eq. (1) showed that no correction was applied if both t_d and t_{d+1} were equal to 24 h. The correction was tested with data from stations that reported hourly data. The hourly precipitation data were available for only 10 weather stations (Table 1).

2.2. Data reconstruction methods

2.2.1. Outline of the reconstruction method application and comparison

Application and comparison of the three methods in this work is illustrated in Fig. 1. Daily precipitation data from all but target weather stations are used as inputs in all three methods. The daily precipitation at the target weather station is the output from all three methods.

The application of the artificial neural network (ANN) as shown in panel (a) in Fig. 1 is straightforward. The application of regression tree (RT) is also straightforward (panel (b) in Fig. 1). As shown in this panel, the RT works not only as a predictive tool but also as the classifier of inputs by their relative influence. This feature of RT is used in the proposed method as illustrated in the panel (c) in Fig. 1. RT is used only as the input classifier. Only data from the most influential weather stations are used as inputs for the ANN to generate predictions for the target station.

The bootstrap resampling method was used for (1) selection of the most influential neighboring stations and (2) statistical evaluation of the reconstruction accuracy for with all three methods. Details on applied ANN, RT, their proposed combination RT + ANN, and resampling are presented below in Sections 2.2.2-2.3.1.

2.2.2. Artificial neural network (ANN) modeling

The feed-forward backpropagation network implementation in Matlab was used in this study (MatlabNeural Network Toolbox 6, Demuth et al., 2008). The feed-forward network had one hidden layer where activation function were logistic sigmoid ('logsig' in Matlab), followed by an output layer where activation function was linear ('poslin' in Matlab) (Demuth et al., 2008). Therefore the equation for the threshold unit output was:

$$y = \mathrm{poslin}(\mathbf{W}_2[\mathrm{logsig}(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1)] + \mathbf{b}_2) \quad (2)$$

where $\mathbf{x}$ is the input data vector, y is the target value, $\mathbf{W}_1$ and $\mathbf{W}_2$ are the weight matrices for the hidden and the output layer, respectively, and $\mathbf{b}_1$ and $\mathbf{b}_2$ are the bias vectors for the hidden and the output layer, respectively. The number of neurons in the hidden layer was defined as $\sqrt{n_{in} \cdot n_{out}}$ where n_{in} and n_{out} are the number of input and output variables, respectively (Maren et al., 1990). The value of n_{out} was one, and the value of n_{in} was the total number of stations less one in the case of ANN application shown in panel (a) in Fig. 1. In the case of ANN application shown in the panel (c) in Fig. 1, the value of n_{in} was the number of the most influential stations resulted from RT. The ANNs were trained using the Levenberg–Marquardt algorithm ('trainlm' in Matlab) to minimize root mean square error between the actual and simulated daily

Table 1
Selected data from the weather stations used in this study.

Station name	Station symbol	Station #	Latitude (°)	Longitude (°)	Elevation (m)	Percentage of valid daily precipitation data (%)[a]	Average annual precipitation (mm)	Average probability of non-zero daily precipitation (%)	Ending integration time fordaily precipitation record (h)
Cumberland 2	NW1	182282	39.65	−78.75	222.5	99.8 (1)	983.4	59.7	7, 18[c]
Frostburg 2	NW2	183415	39.67	−78.93	661.4	99.5 (5)	1126.6	60.0	7
Altoona 3W	NW3	360140	40.50	−78.47	402.3	98.7 (2)	1121.3	46.3	9, 24[c]
Buckstown 1 SE	NW4	361073	40.07	−78.85	749.8	98.7 (8)	951.6	64.8	7
Dunlo	NW5	362298[b]	40.28	−78.72	719.3	98.5 (30)	1107.4	59.6	7
Ebensburg Sewage Plt	NW6	362470	40.47	−78.73	591.3	98.6 (30)	1191.1	67.0	8, 16[c]
Everett	NW7	362721[b]	40.02	−78.37	304.8	99.8 (1)	982.4	54.2	7,9[c]
Laurel Summit	NW8	364839	40.17	−79.13	832.1	98.6 (31)	1450.8	64.6	6,7[c]
Meyersdale 2 SSW	NW9	365686	39.78	−79.03	609.6	98.7 (31)	1027.1	61.7	7
Raystown Lake 2	NW10	367312[b]	40.38	−78.08	304.8	98.8 (3)	979.9	54.6	8
Saxton 1w	NW11	367846	40.12	−78.15	253.0	99.8 (1)	1054.2	51.1	7
Shawnee Sp	NW12	368011	40.03	−78.63	388.6	99.4 (2)	990.6	57.0	8
Somerset	NW13	368244	40.00	−79.08	640.1	99.5 (4)	1199.7	68.9	8
Tyrone	NW14	369022[b]	40.67	−78.22	271.3	98.6 (31)	1087.8	55.9	7
Williamsburg	NW15	369714	40.47	−78.20	257.6	100.0 (1)	1046.9	56.6	7
Wolfsburg	NW16	369823	40.05	−78.53	361.2	100.0 (1)	998.4	56.8	7
Millers 4 NE	NE1	185934[b]	39.72	−76.80	262.1	100.0 (0)	1164.8	54.8	18
Biglerville	NE2	360656	39.93	−77.25	219.5	99.7 (2)	1095.6	51.4	7
Chambersburg 1 ESE	NE3	361354	39.93	−77.63	195.1	100.0 (1)	1101.6	48.9	9,21[c]
Dehart Dam	NE4	362071	40.47	−76.87	163.7	99.6 (3)	1257.9	52.4	7
Hanover 4 SW	NE5	363665	39.77	−77.03	170.1	100.0 (1)	1108.1	52.1	8
Harrisburg 1 NE	NE6	363698	40.28	−76.87	128.0	99.8 (2)	1134.1	51.2	7
Lewistown	NE7	364992	40.58	−77.57	140.2	99.8 (1)	1066.0	52.9	8, 20[c]
Middletown Harrisburg International Airp.	NE8	365703[b]	40.20	−76.77	95.1	99.9 (1)	1089.7	34.3	24
Pine Grove Furnace	NE9	366955	40.03	−77.30	269.7	98.3 (3)	1113.7	50.4	8
Shippensburg	NE10	368073	40.05	−77.52	207.3	100.0 (1)	1054.7	33.0	24
South Mountain	NE11	368308	39.85	−77.48	463.3	99.9 (1)	1165.0	52.2	7
Spring Grove	NE12	368379	39.87	−76.87	137.2	99.9 (1)	1096.2	49.5	8
York 3 Ssw Pump Station	NE13	369933	39.92	−76.75	118.9	100.0 (1)	1153.5	51.8	17,24[c]
York Haven	NE14	369950	40.12	−76.72	94.5	98.3 (31)	1087.5	49.5	7
Potomac Fltr Plt	S1	187272	39.03	−77.25	82.3	99.1 (2)	1108.1	48.4	8
Front Royal	S2	443229	38.90	−78.18	283.5	99.0 (4)	1080.0	46.4	7, 24[c]
Star Tannery	S3	448046[b]	39.08	−78.45	289.6	99.3 (7)	942.9	38.8	7
Sterling (res)	S4	448084	38.98	−77.48	85.3	100.0 (0)	1155.1	50.6	7
Washington Dc Dulles Airp.	S5	448903[b]	38.93	−77.47	88.4	99.8 (3)	1068.2	32.8	24
Winchester	S6	449181	39.18	−78.15	219.5	99.8 (1)	1021.3	51.7	7
Winchester 7 SE	S7	449186	39.18	−78.12	207.3	98.5 (31)	990.3	48.7	7,8[c]
Martinsburg E WvRgnlAirp.	S8	465707[b]	39.40	−77.98	162.8	100.0 (1)	974.3	33.6	24
Romney 1 SW	S9	467730[b]	39.33	−78.77	219.5	98.7 (2)	934.6	50.2	7,8[c]

[a] Number in parenthesis is the longest number of consecutive missing days.
[b] Hourly precipitation data are available.
[c] Ending integration times varied between January 1, 2001 and December31, 2007.

precipitation at the target weather station. Initial weights matrices and bias vectors were randomly generated from a uniform distribution (Demuth et al., 2008).

2.2.3. Regression tree (RT) modeling

The regression tree (RT) modeling can be viewed as an exploratory technique based on uncovering structure in data (Clark and Pregibon, 1991). The RT model implemented in this work was coded in Matlab and tested following Rawls and Pachepsky (2002) work where the RT algorithm was described in detail and a test example was given. The original data were partitioned first into two groups, then into four groups, and so forth, providing groups as homogeneous as possible at each level of partitioning. The homogeneity of a dataset was measured by its deviance D, described by the sum of squared differences between the individual response values (y_i) and average response values ($\bar{y}$) over the dataset:

$$D = \sum_i (y_i - \bar{y})^2 \tag{3}$$

Non-homogeneity minimization was achieved by maximizing ΔD:

$$\Delta D = D - D_{s1} - D_{s2} \tag{4}$$

where D equals the deviance of the dataset before partition, and D_{s1} and D_{s2} are the deviances of the two subsets formed by the partition. The subsets were further partitioned into two until further partitioning did not provide positive values of ΔD or the total number of partitions became greater than 10. The average value of the response variable for the terminal partitions served as the regression tree-predicted value. For each station, the neighboring stations that provided the partitions were recorded as the influential neighboring stations (Fig. 1, panels (b) and (c)). The neighboring station that provided the first partition of the target station was regarded as the most influential neighboring station.

2.2.4. Artificial neural network after regression tree (RT + ANN)

A two-step method, RT + ANN, which consisted of determining the influential predictors with RT and reconstructing the precipitation with ANN using only influential predictors as inputs was used in this work as shown in Fig. 1, panel (c). Properties of RT and ANN were the same as their separate applications (Fig. 1, panels (a) and (b)). This method is conceptually similar to the approach used by Faucher et al. (1999) to impute data on marine winds. It was hypothesized that the RT + ANN data reconstruction would be more robust than the ANN-only or RT-only reconstruction, and this

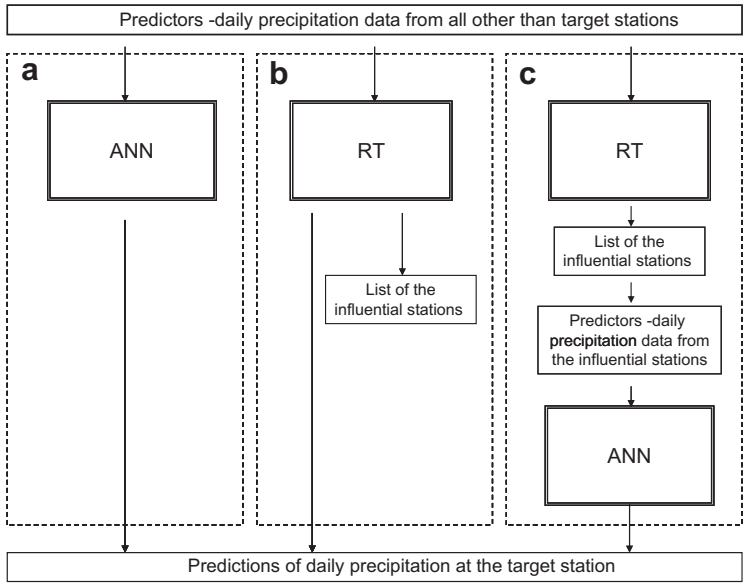

Fig. 1. Three missing precipitation data reconstruction methods applied in this work: (a) the feed-forward artificial neural network (ANN) with one hidden layer, (b) regression tree (RT), (c) regression tree to select the influential input stations and to use only data from these stations in the feed-forward artificial neural network with one hidden layer (RT + ANN).

hypothesis was tested by evaluating the uncertainty of the data reconstruction as described in the following section.

2.3. Evaluation of the reconstruction methods

2.3.1. Bootstrap evaluation of the reconstruction accuracy

The accuracy of the reconstruction was characterized with the Pearson correlation coefficient (r) and root mean square error (RMSE). The bootstrap method was applied to evaluate the uncertainty in values of these accuracy measures. For the bootstrapping (Good, 1999), the original dataset was randomly sampled with a replacement, and 90% and 10% of the bootstrapped dataset were used for training and testing, respectively, in the application of reconstruction methods. Once the training was finished, the trained model was applied to the testing dataset, and the accuracy of the reconstruction was assessed in terms of the correlation coefficient r and RMSE. The bootstrapping was repeated 1000 times in order to estimate probability distribution functions of r and RMSE.

The probability distribution functions of r and RMSE were compared using the Kolmogorov–Smirnov (KS) test to assess the significance of differences in the performances of the three reconstruction methods. The two-sample KS test subroutine ('kstest2.m') of Matlab statistics toolbox (MathWorks, 2009) was used to compute the KS statistic between the performances of data reconstruction methods.

Bootstrap results were also used to develop the complete list of influential stations. All stations that were recorded as influential in any of bootstrap runs were included in the list.

2.3.2. Functional evaluation of the reconstruction methods

Since the ultimate goal of the precipitation data reconstruction was to use the reconstructed precipitation data in the SWAT model, the reconstruction error propagation in SWAT was investigated. The 1-year (2007) precipitation data time series were generated for three stations with the three reconstruction methods introduced above. SWAT simulations of streamflow were conducted using both the generated precipitation data and the observed precipitation data for the same date.

The SWAT model was used within the BASINS environmental analysis system (USEPA, 2001). For three weather stations, the smallest watersheds including these weather stations were delineated with the 'Automatic Watershed Delineation' tool in BASINS using the Digital Elevation Model (DEM). The land-use [Land Use/Land Cover (LULC) data of United States Geological Survey (USGS)] and soil [State Soil Geographic (STATSGO) database of United States Department of Agriculture/Natural Resources Conservation Service (USDA/NRCS)] maps were employed by 'Land Use, Soil Classification and Overlay' tool in BASINS. All hydraulic input parameters, such as the main channel Manning's roughness coefficient, "n", rainfall/runoff/routing option, potential evapotranspiration method, channel water routing method, surface runoff lag coefficient, reach evaporation adjustment factor, were set to default values in SWAT.

The daily streamflow values at the outlets of each subwatershed simulated with the actual and the reconstructed precipitation were compared in terms of the r.

3. Results and discussion

3.1. Preliminary precipitation data analysis

The 39 weather stations within the study area had mostly scattered missing precipitation data for 7 years from 2001 to 2007 except stations NW5, NW6, NW8, NW9, NW14, NE14, and S7 having a month or more of consecutive missing daily values (Table 1). Overall, more than 98% of valid (non-missing) daily precipitation data was available. Over the 7-year observation period, the average annual precipitation varied from 934 to 1450 mm across the study area. The average annual probability of a non-zero daily precipitation varied from 33% to 69% (Table 1).

The stations had different ending integration times between 2001 and 2007 (Table 1); some accumulated daily data from 7 am to the next 7 am, some from 8 am to the next 8 am, some from 6 pm to the next 6 pm, and some from 12 am to 12 am. As expected, daily precipitation data that were accumulated by the ending of the integration time other than midnight did not correlate well with the daily precipitation computed based on the hourly precipitation integration over 24 h counting from midnight. The longer the time period between the ending of the integration time

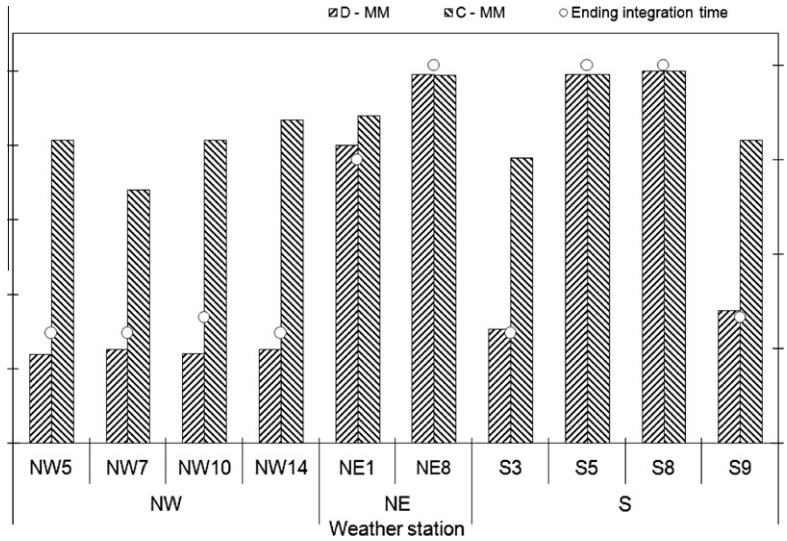

Fig. 2. Correlations between reported daily precipitation (D) and integrated 'midnight-to-midnight' hourly precipitation (MM) (▨: D – MM) and between daily precipitation corrected by Eq. (1) (C) and MM (▧: C – MM), and ending integration times of D (●○).

and midnight, the worse was the correlation (Fig. 2). Data in Fig. 2 shows the correction according to Eq. (1) had substantially improved the daily precipitation data quality. The corrected daily precipitation dataset (39 stations and 7 years) was sorted by calendar date, and data were removed for the days when any of the stations had a missing value to evaluate the reconstruction methods. Finally, a total of 1816 days were available, and thus the randomly chosen 1634-days data (90%) and 182-days data (10%) were used for training and testing, respectively, in each bootstrap run.

3.2. Influential neighbors from the regression tree analysis

Fig. 3 shows the most influential neighbors for each station found from the RT analysis. The thickness of arrows in this figure

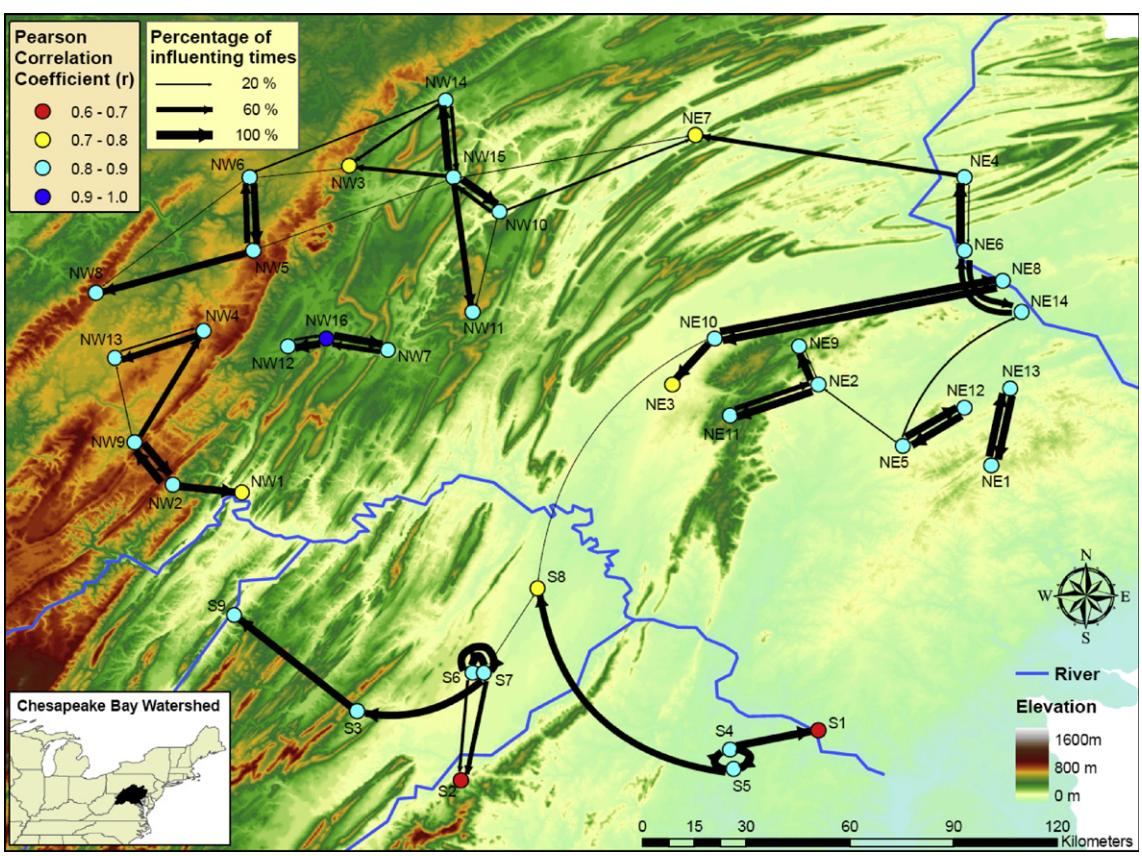

Fig. 3. Similarity between precipitation time series recorded at weather stations in the study area; the thickness of arrows indicates the percentage of times in the regression tree (RT) bootstrap when the station at the arrow nock was the most influential neighbor for the station at the arrow head. The fill colors in station symbols show the Pearson correlation coefficient (r) between actual and reconstructed with RT daily precipitation.

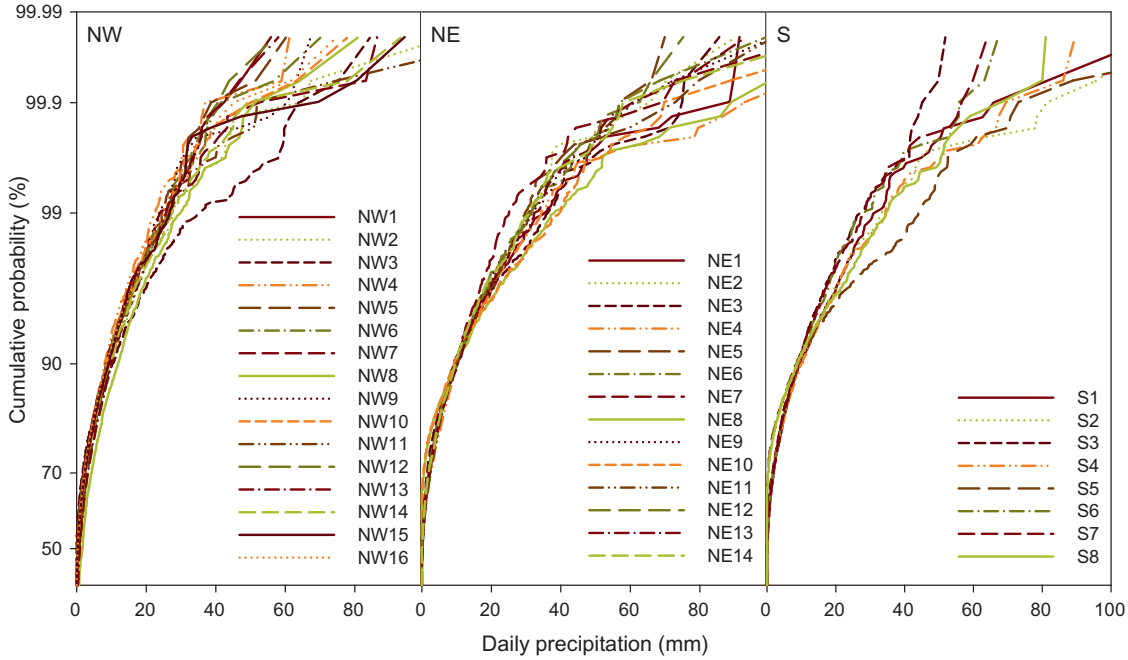

Fig. 4. Cumulative probability distribution functions of the corrected daily precipitation at the weather stations in the study area.

shows the percentage of bootstrap runs in which the station at the arrow nock has been the most influential in RT for the station at the arrow head, i.e. has provided the first split in the RT. Data in Fig. 3 shows that precipitation time series at weather stations in the Appalachian Mountains were dissimilar to the time series observed at the coastal plain in the eastern part of the study area. Weather stations to the south of Potomac River formed a separate group. Accordingly, the study area was divided into three subareas – mountainous (NW, northwest), coastal plain (NE, northeast), and south (S). The cumulative probability distributions of daily precipitation in the three subareas are shown in Fig. 4. The average annual precipitation in the NE subarea (1120.6 mm) was similar to that in the NW subarea (1081.2 mm) ($p > 0.2$), but larger than that in S subarea (1030.5 mm) ($p < 0.009$). The average probability of non-zero daily precipitation was higher ($p < 0.0003$) in the mountainous NW subarea than in rather plain NE and S subareas.

The most influential stations were the closest ones in 21 of 39 cases. In the mountainous NW subarea, the selection of the most influential stations reflected the landscape position. For example, the station NW5 was the most influential one for station NW8 which was located at a similar elevation. The station NW13 was closer to station NW8 than the station NW5, but was located in a valley between ridges and was not influential for the station NW8. Likewise, the station NW4 had the most influence on the station NW9 rather than on the station NW12; the latter was closer to station NW4 but was located in a piedmont valley. These results concurred with earlier observations of the effect of the landscape position on the precipitation distributions in mountainous areas (e.g., Daly et al., 1994; Sevruk, 1997).

In the NE subarea, the selection of the most influential station was the result of the apparent predominant precipitation connectivity in the East–West direction. For example, the most influential neighbor of the station NE8 was the station NE10 rather than two very close stations NE6 and NE14. The station NE2 had the station NE11 as the most influential neighbor but not the closer station NE9. The station NE5 rather than the station NE13 was selected as the most influential neighbor for the station NE12.

In the S subarea, most of stations had the closest station as their most influential neighbors, station S8 being an exemption. This station along with station S5 had a very low number of rainy days (Table 1) and that may explain selection of the S5 as the influential neighbor.

Median values of r between RT predictions and measured daily precipitation were in the range from 0.8 to 1.0 for 33 of 39 stations (Fig. 3). Some of poorer correlations could be explained by the fine-scale variation in the probability of non-zero daily precipitation which was quite substantial in the study area. For example, the distance between stations NE8 and NE14 was only 10 km, and yet the average probabilities of non-zero daily precipitation at NE8 and NE14 were 34% and 50%, respectively (Table 1). The station NW3 had the average probability of non-zero precipitation equal to 46.3% whereas all its neighbors had probability above 55%. The station S8 was relatively far from other weather stations except stations S6 and S7. The average probability of non-zero precipitation at the station S8 (33.6%) was significantly different from those of stations S6 (51.7%) and S7 (48.7%), but similar to that of the station S5 (32.8%) which, as a result, served as the most influential neighbor of the station S8 (Fig. 3). The poor performance of the precipitation reconstruction for the station S8 was presumably due to the generally low correlation with its closest neighbors and its remoteness from other weather stations. In case of the station NE7, the low r values were probably related to the large distance to its neighbors. Overall, most weather stations with low r values were never influential for any other weather stations, and were influenced by more than one neighbor. One more reason for low reconstruction accuracy could be the shape of the cumulative probability distribution of daily precipitation. For example, the station NW3 had substantially different cumulative probability distribution as compared with all other stations in the subarea (Fig. 4-NW).

3.3. Performance of reconstruction methods

The performance of the three methods, i.e. RT, ANN, and RT + ANN, in reconstruction of missing precipitation data is compared in terms of r and RMSE using the KS probability values in Table 2. The median values and the 90% confidence intervals of r and RMSE for stations grouped by subareas are shown in

Table 2
Number of influential stations, the median Pearson correlation coefficient (*r*), the median root mean square error (RMSE, mm), and the Kolmogorov–Smirnov (KS) test probabilities that there is no difference between the distributions of *r* (KS_*r*) and distributions of RMSE (KS_RMSE) for three precipitation reconstruction methods – regression tree (RT), artificial neural network (ANN), and ANN with inputs selected after RT analysis (RT + ANN). Statistics were computed from 1000 bootstrap runs for each station.

Station	# of influ. Stns.	Median *r*			KS_*r* (*p*-value)			Median RMSE (mm)			KS_RMSE (*p*-value)		
		RT	ANN	RT + ANN	RT vs. ANN	RT *vs.* RT + ANN	ANN *vs.* RT + ANN	RT	ANN	RT + ANN	RT vs. ANN	RT *vs.* RT + ANN	ANN *vs.* RT + ANN
NW1	1	0.703	0.708	**0.748**	0.000	0.000	0.000	3.66	3.72	**3.39**	0.032	0.000	0.000
NW2	1	0.873	0.879	**0.891**	0.000	0.000	0.000	2.81	2.73	**2.55**	0.000	0.000	0.000
NW3	3	0.750	0.739	**0.822**	0.006	0.000	0.000	4.58	4.80	**3.87**	0.001	0.000	0.000
NW4	2	0.840	0.857	**0.860**	0.000	0.000	0.003	2.69	2.55	**2.45**	0.000	0.000	0.000
NW5	2	0.822	0.841	**0.862**	0.000	0.000	0.000	2.95	2.81	**2.59**	0.000	0.000	0.000
NW6	2	0.833	0.854	**0.875**	0.000	0.000	0.000	2.95	2.81	**2.55**	0.000	0.000	0.000
NW7	1	0.887	0.843	**0.914**	0.000	0.000	0.000	2.37	2.98	**2.08**	0.000	0.000	0.000
NW8	2	0.823	0.821	**0.843**	0.000	0.000	0.000	3.52	3.49	**3.30**	0.003	0.000	0.000
NW9	1	0.874	0.879	**0.895**	0.000	0.000	0.000	2.61	2.51	**2.32**	0.007	0.000	0.000
NW10	1	0.871	0.887	**0.907**	0.000	0.000	0.000	2.49	2.40	**2.18**	0.001	0.000	0.000
NW11	2	0.837	0.806	**0.866**	0.000	0.000	0.000	2.89	3.29	**2.69**	0.000	0.000	0.000
NW12	1	0.889	0.882	**0.919**	0.000	0.000	0.000	2.35	2.44	**2.03**	0.016	0.000	0.000
NW13	2	0.869	0.875	**0.885**	0.000	0.000	0.000	2.71	2.65	**2.48**	0.001	0.000	0.000
NW14	1	0.848	0.847	**0.874**	0.000	0.000	0.000	3.03	3.11	**2.73**	0.019	0.000	0.000
NW15	2	0.886	0.902	**0.925**	0.000	0.000	0.000	2.49	2.36	**1.99**	0.000	0.000	0.000
NW16	2	0.920	0.928	**0.952**	0.000	0.000	0.000	1.93	1.84	**1.51**	0.000	0.000	0.000
NE1	1	0.847	0.811	**0.868**	0.000	0.000	0.000	3.58	4.19	**3.45**	0.000	0.000	0.000
NE2	3	0.863	0.877	**0.922**	0.000	0.000	0.000	3.08	3.09	**2.36**	0.000	0.000	0.000
NE3	1	0.737	0.697	**0.784**	0.000	0.000	0.000	5.02	5.36	**4.50**	0.000	0.000	0.000
NE4	1	0.861	0.824	**0.874**	0.000	0.001	0.000	3.50	4.04	**3.31**	0.000	0.000	0.000
NE5	1	0.876	0.874	**0.891**	0.001	0.000	0.000	3.07	3.20	**3.00**	0.000	0.000	0.000
NE6	2	0.896	0.868	**0.906**	0.000	0.000	0.000	2.63	2.98	**2.52**	0.000	0.000	0.000
NE7	3	0.736	0.711	**0.793**	0.000	0.000	0.000	3.85	4.12	**3.41**	0.000	0.000	0.000
NE8	1	0.808	0.786	**0.810**	0.000	0.214	0.000	4.26	4.63	**4.10**	0.000	0.001	0.000
NE9	1	0.849	0.844	**0.868**	0.001	0.000	0.000	3.25	3.34	**3.04**	0.000	0.000	0.000
NE10	1	**0.851**	0.781	0.810	0.000	0.000	0.000	**3.77**	4.52	4.16	0.000	0.000	0.000
NE11	1	0.870	**0.880**	0.877	0.000	0.000	0.565	3.15	3.04	**3.01**	0.002	0.000	0.004
NE12	1	0.879	0.868	**0.901**	0.001	0.000	0.000	2.92	3.08	**2.70**	0.000	0.000	0.000
NE13	1	0.854	0.847	**0.878**	0.036	0.000	0.000	3.15	3.36	**2.99**	0.000	0.000	0.000
NE14	2	0.856	0.870	**0.883**	0.000	0.000	0.000	3.17	3.02	**2.88**	0.000	0.000	0.000
S1	1	0.644	0.572	**0.670**	0.000	0.000	0.000	5.19	5.60	**4.89**	0.000	0.000	0.000
S2	2	0.695	0.615	**0.734**	0.000	0.000	0.000	4.76	5.44	**4.43**	0.000	0.000	0.000
S3	1	0.819	0.810	**0.842**	0.000	0.000	0.000	3.16	3.21	**2.92**	0.000	0.000	0.000
S4	1	**0.862**	0.827	0.816	0.000	0.000	0.000	**3.45**	3.97	3.95	0.000	0.000	0.016
S5	1	0.801	0.776	**0.811**	0.000	0.019	0.000	4.80	5.09	**4.50**	0.000	0.000	0.000
S6	1	0.867	0.850	**0.891**	0.000	0.000	0.000	2.80	3.02	**2.58**	0.000	0.000	0.000
S7	1	0.862	0.867	**0.889**	0.000	0.000	0.000	2.90	2.79	**2.60**	0.000	0.000	0.000
S8	3	0.737	0.716	**0.823**	0.000	0.000	0.000	4.87	5.04	**3.93**	0.001	0.000	0.000
S9	1	**0.817**	0.810	0.777	0.005	0.000	0.000	**3.08**	3.12	3.33	0.014	0.000	0.000

Numbers in bold indicate the best results among three methods.

Fig. 5. Overall, the accuracy of RT estimates was comparable with the accuracy of ANN. The median values of *r* were 0.831 and 0.819, and RMSE were 3.3 mm and 3.5 mm for RT and ANN, respectively. Table 2 shows that RT provided better estimates in cases when the predictions were generally less accurate with both methods, e.g. for stations NE3, S1, and S2 (Table 2). The confidence intervals for the *r* values (Fig. 5) in the RT method were substantially narrower when compared with those in the ANN method, i.e. the RT method provided more consistent, or robust, reconstruction results than ANN method in this study.

The numbers of the most influential neighbors encountered in the bootstrap runs for each weather station varied from 1 to 3 as shown in Table 2. Using the data only from the RT-derived most influential neighbors as ANN inputs led to the increase in accuracy and robustness as compared to either RT or ANN alone. Pearson correlation coefficients for RT + ANN method were higher and the RMSE values were lower for all stations except the stations NE10, S4, and S9 (Table 2). Overall, the median *r* and RMSE of RT + ANN were 0.855 and 3.1 mm, respectively. The *p*-values with less than 0.05 in KS tests (Table 2) supported the superiority of RT + ANN method. Interestingly, the highest accuracy was achieved in the mountainous NW subarea where the average probability of the non-zero daily precipitation was higher than in other subareas (Table 1 and Fig. 5).

It was interesting for reference purposes to compare the accuracy of precipitation reconstruction in this work with results obtained in other reconstruction research. Such comparisons showed that comparable or even better accuracy had been attained in this study. Kuligowski and Barros (1998) applied the backpropagation neural network to reconstruct the daily precipitation and their *r* values were in the range from 0.458 to 0.823. Coulibaly and Evora (2007) investigated six different types of ANN and provided *r* ranging from 0.575 to 0.869. The RT + ANN method in this study resulted in values of *r* ranging from 0.670 to 0.952 (from 0.748 to 0.952 in NW subarea, from 0.784 to 0.922 in the NE subarea, and from 0.670 to 0.891 in the S subarea). The performance of RT + ANN method in the mountainous NW region (RMSE = 2.54 mm) was comparable to the Abebe et al. (2000) results from the application of multi-layer perceptron ANN with inputs from two neighboring stations within 15 km in Italian mountain area (RMSE = 2.63 mm). All three reconstruction methods in this study resulted in better performance than the Group Method of Data Handling with the data for the same station on the dates when the data was available (Acock and Pachepsky, 2000; RMSE = 6.5 mm) and the 'Closest Station' method with cluster analysis (Garcia et al., 2006; RSME = 9.05 mm).

The uncertainty in Fig. 5 predictions is relatively large even with the RT + ANN method which had the smallest uncertainty among

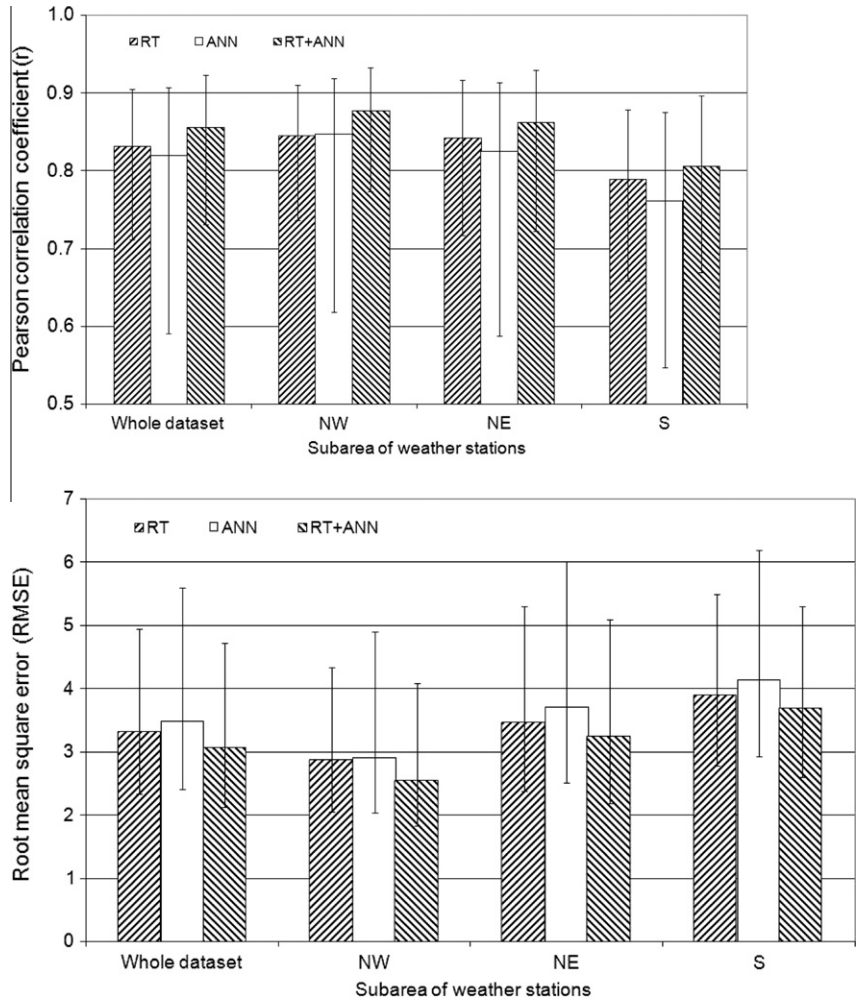

Fig. 5. Accuracy of the daily precipitation reconstruction with regression tree (RT), artificial neural network (ANN), and ANN with inputs selected after RT analysis (RT + ANN) in terms of the median Pearson correlation coefficient (r) and the root mean square error (RMSE, mm) with 90% confidence intervals from bootstrap runs. Results were pooled for the whole dataset and for three subareas – northwest (NW), northeast (NE), and south (S).

three reconstruction methods. The weather data imputation by no means is meant to replace measurements. The most common purpose of the imputation is to provide an uninterrupted data series as required by simulation models. The precondition for such an application is the small percentage of missing data. If the data imputation is applied, an error propagation study seems to be in order to establish the effect of the imputation errors on the output variables of the model in question. Other imputation strategies such as using precipitation climatology can be tested. It is quite possible that additional environmental variables may be useful for imputation in a specific geographical situation. Machine-learning methods are not parameter-free and changing these parameters, e.g. training-to-testing proportion, can affect in the accuracy. Other machine-learning methods, e.g. distributed support vector regression (Wu et al., 2008) or model trees (Solomatine and Xue, 2004), can be efficient in data imputation. The performance of ANN can be further improved using ensemble techniques with preliminary data preprocessing using canonical correlation analysis (Shu and Burn, 2004; Shu and Ouarda, 2007). Results of this work underscore the appropriateness of the preliminary screening of inputs to improve the accuracy and the reliability of data imputation. They also show that the accuracy of the data infilling method is site-specific, and this is evidenced by the differences between the accuracy in the three subareas in this study.

3.4. Performance of reconstruction methods as assessed with SWAT streamflow simulations

The functional evaluation of the precipitation reconstruction was done for three watersheds (Fig. 6) that included one of weather stations within each subarea. Stations NW16 (#369823), NE2 (#360656), and S6 (#449181) that had the maximum median r in each subarea (Table 2) were selected for the SWAT streamflow simulations for 1-year (2007). Sub-watersheds of stations NE2 (114 km^2) and S6 (76 km^2) had first-order streams, and the subwatershed of the station NW16 (379 km^2) included two orders of streams and was relatively larger than other two as shown in Fig. 6. The main channel lengths were 35 km, 20 km, and 21 km, the main channel depths were 1.2 m, 0.9 m, and 0.7 m, the main channel slopes were 0.003 m m^{-1}, 0.005 m m^{-1}, and 0.000 m m^{-1} (all slopes under the spatial resolution of the DEM of 75 m), and the main channel widths were 35 m, 22 m, and 17 m for the sub-watersheds of the stations NW16, NE2, and S6, respectively. According to the USGS classification, the predominant landuse class in the subwatershed of station NW16 was 'Forest – Deciduous', and those in the subwatershed of stations NE2 and S6 were 'Agricultural Land – Generic'. Sub-watersheds of stations NW16 and S6 were dominated by loamy soils of the PA033 soil association and by clay soils of the VA069 soil association, whereas silt

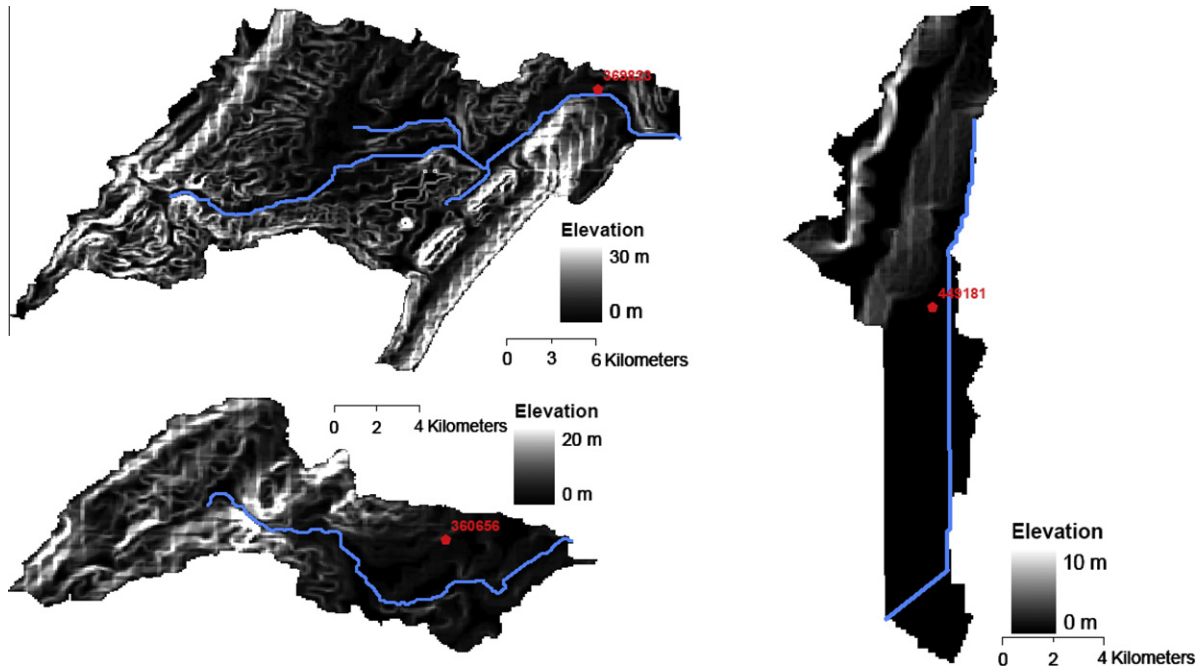

Fig. 6. Subwatersheds for weather stations NW16 (#369823), NE2 (#360656), and S6 (#449181) used in streamflow simulations with the Soil and Water Assessment Tool (SWAT). Symbols with the cooperative station ID number of the National Weather Service show locations of weather stations.

loam soils of PA063 and PA064 soil association and sandy loam soils of the PA066 soil association were almost equally represented in the subwatershed of station NE2. The streamflow simulations were conducted using the curve number runoff option and the variable storage method for channel water routing. The predominant runoff curve numbers were 91, 86 and 86 for the sub-watersheds with stations NW16, NE2, and S6, respectively, determined based on each predominant landuse and soil. The surface runoff lag coefficient was set to 4.0 days, the Priestley–Taylor method for potential evapotranspiration was applied with the reach evaporation adjustment factor = 1.0 and the leaf area index at which no evaporation occurs from water surface equal to 3.0 $m^2 \cdot m^{-2}$. The main channel Manning's roughness coefficient, "*n*", was set to 0.014.

The accuracy of precipitation predictions and streamflow predictions was compared for stations NW16, NE2 and S6. For this specific case of 2007, values of *r* for the precipitation reconstruction using RT and ANN methods were mostly smaller than their median values in Table 2 at all three stations while the RT + ANN method gave values of *r* comparable with median values.

In the subwatershed of the station NW16, the accuracy of streamflow predictions in terms of *r* was substantially lower than that of precipitation estimates with both RT and ANN stand-alone methods (Fig. 7). In simpler and smaller sub-watersheds (stations NE2 and S6), the accuracy of streamflow predictions was higher than the accuracy of precipitation estimates when the ANN was used to reconstruct the data. The accuracy of streamflow

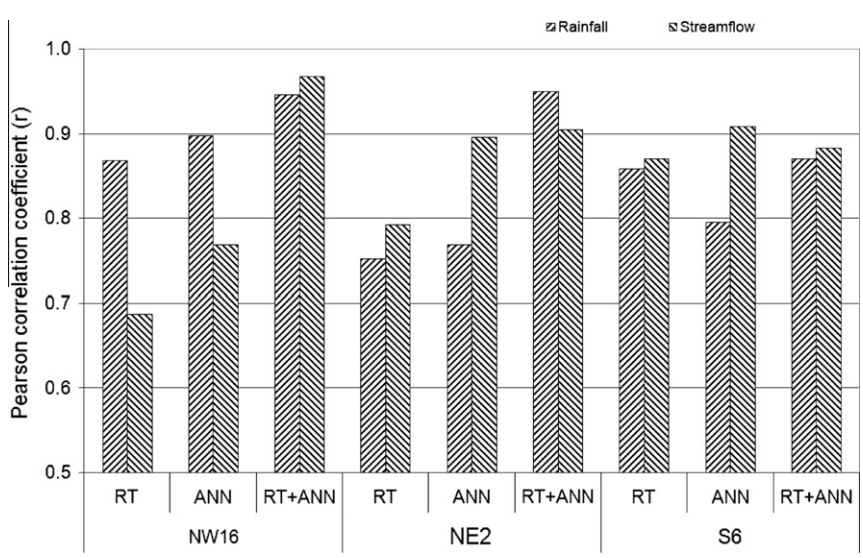

Fig. 7. Pearson correlation coefficients (*r*) between reconstructed and actual precipitation and between streamflow simulated with SWAT model using reconstructed and actual precipitation for three weather stations in the northwest (NW), northeast (NE), and South (S) subareas. Precipitation was reconstructed with the regression tree (RT), the artificial neural network (ANN) and ANN with inputs selected after RT analysis (RT + ANN).

predictions was similar to the accuracy of the precipitation prediction for all three sub-watersheds when the two-step RT + ANN method was used. In this evaluation, the RT + ANN method appeared to be the most robust precipitation reconstruction method for the purposes of streamflow simulations.

4. Conclusion

Two machine-learning methods, regression trees (RT) and artificial neural networks (ANN), were used alone and then in combination (RT + ANN) to reconstruct missing precipitation data using daily precipitation time series from 39 weather stations in the central part of Chesapeake Bay Watershed. The ANN reconstruction employed data from all stations as inputs whereas the RT + ANN method built ANN with inputs only from the influential stations found from the preliminary bootstrap RT application. The performance of the methods was compared using the Pearson correlation coefficients and the root mean square error. The accuracy of streamflow simulations with the SWAT model using the reconstructed precipitation data was also utilized to compare the performance of the three reconstruction methods. Bootstrap was used with all three methods to assess the uncertainty in the performance measures.

The RT provided a transparent visual representation of the similarity between the stations in their daily precipitation time series. The nearest station in terms of the distance was the most influential RT predictor only for about half of all stations. This could be related to topography in the mountainous subarea and to the apparent fine-scale variations in coastal plain subarea. For each station, from one to three stations could be found that were the most influential predictors in bootstrapped RT runs. Both RT and ANN provided reconstruction accuracy comparable to (or better than) other published results of ANN application to precipitation reconstruction. The proposed RT + ANN method significantly improved accuracy and was more robust when compared with RT and ANN alone. This method was also more accurate and robust in streamflow predictions with reconstructed precipitation.

Although the newly suggested method in this study, RT + ANN, showed a better performance in infilling precipitation data, the uncertainty of results was relatively large. Using both other methods of machine learning and other set of predictors can be beneficial. This study demonstrated the appropriateness of the preliminary screening of inputs to improve the accuracy and the reliability of data imputation, and the value of functional evaluation of the data reconstruction method. It is concluded that the type of analysis described in this work can be useful in the precipitation reconstruction and warrants research on its applicability to other weather variables.

References

Abebe, A.J., Solomatine, D.P., Venneker, R.G.W., 2000. Application of adaptive fuzzy rule-based models for reconstruction of missing precipitation events. Hydrological Sciences Journal 45 (3), 425–436.

Acock, M.C., Pachepsky, Y.A., 2000. Estimating missing weather data for agricultural simulations using group method of data handling. Journal of Applied Meteorology 39 (7), 1176–1184.

Burrows, W.R., Benjamin, M., Beauchamp, S., Lord, E.R., McCollor, D., Thomson, B., 1995. CART decision-tree statistical-analysis and prediction of summer season maximum surface ozone for the Vancouver, Montreal, and Atlantic regions of Canada. Journal of Applied Meteorology 34 (8), 1848–1862.

Clark, L.A., Pregibon, D., 1991. Tree-based models. In: Chambers, J.M., Hastie, T.J. (Eds.), Statistical Models in S. Wadsworth. Pacific Grove, CA, pp. 377–419.

Coulibaly, P., Evora, N.D., 2007. Comparison of neural network methods for infilling missing daily weather records. Journal of Hydrology 341 (1–2), 27–41.

Daly, C., Neilson, R.P., Phillips, D.L., 1994. A statistical topographic model for mapping climatological precipitation over mountainous terrain. Journal of Applied Meteorology 33 (2), 140–158.

Demuth, H., Beale, M., Hagan, M., 2008. Matlab: Neural Network Toolbox 6 – User's Guide. The MathWorks.

Faucher, M., Burrows, W.R., Pandolfo, L., 1999. Empirical-statistical reconstruction of surface marine winds along the western coast of Canada. Climate Research 11 (3), 173–190.

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

Franklin, M., Kotamarthi, V.R., Stein, M.L., Cook, D.R., 2008. Generating data ensembles over a model grid from sparse climate point measurements. Journal of Physics: Conference Series 125, 012019.

Garcia, B.I.L., Sentelhas, P.C., Tapia, L., Sparovek, G., 2006. Filling in missing rainfall data in the Andes region of Venezuela, based on a cluster analysis approach. Revista Brasileira Agrometeorologia 14 (2), 225–233.

Good, P.I., 1999. Resampling Methods: A Practical Guide to Data Analysis. Birkhauser Boston, c/o Springer-Verlag, New York, Inc., New York, NY.

Jeffrey, S.J., Carter, J.O., Moodie, K.B., Beswick, A.R., 2001. Using spatial interpolation to construct a comprehensive archive of Australian climate data. Environmental Modelling & Software 16 (4), 309–330.

Kim, J.-W., Pachepsky, Y.A., Shelton, D.R., Coppock, C., 2010. Effect of streambed bacteria release on *E. coli* concentrations: monitoring and modeling with the modified SWAT. Ecological Modelling 221, 1592–1604.

Krysanova, V., Arnold, J.G., 2008. Advances in ecohydrological modelling with SWAT – a review. Hydrological Sciences Journal 53 (5), 939–947.

Kuligowski, R.J., Barros, A.P., 1998. Using artificial neural networks to estimate missing rainfall data. Journal of the American Water Resources Association 34 (6), 1437–1447.

Linacre, E., 1992. Climate Data and Resources – A Reference and Guide. Routledge, London and New York.

Lowry, W.P., 1972. Compendium of Lecture Notes in Climatology for Class IV Meteorological Personnel, Geneva.

Maren, A.J., Harston, C.T., Pap, R.M., 1990. Handbook of Neural Computing Applications. Academic Press, San Diego.

MathWorks, 2009. Matlab: Statistics Toolbox 7 – User's Guide. The MathWorks.

Michaelides, S.C., Tymvyos, F.S., Kalogirou, S.A., 2007. Artificial neural networks for meteorological variables pertained to energy and renewable energy applications. In: Kalogirou, S. (Ed.), Artificial Intelligence in Energy and Renewable Energy Systems. Nova Science Publishers, pp. 47–82.

NCDC, 2008. National Climatic Data Center. <http://www.ncdc.noaa.gov/oa/ncdc.html> (accessed 28.12.2008).

Neitsch, S.L., Arnold, J.G., Kiniry, J.R., Williams, J.R., 2005. Soil and Water Assessment Tool Theoretical Documentation (Version 2005).

Ramos-Calzado, P., Gomez-Camacho, J., Perez-Bernal, F., Pita-Lopez, M.F., 2008. A novel approach to precipitation series completion in climatological datasets: application to Andalusia. International Journal of Climatology 28 (11), 1525–1534.

Rawls, W.J., Pachepsky, Y.A., 2002. Soil consistence and structure as predictors of water retention. Soil Science Society of America Journal 66 (4), 1115–1126.

Santhi, C., Srinivasan, R., Arnold, J.G., Williams, J.R., 2006. A modeling approach to evaluate the impacts of water quality management plans implemented in a watershed in Texas. Environmental Modelling & Software 21 (8), 1141–1157.

Schneider, T., 2001. Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. Journal of Climate 14 (5), 853–871.

Sevruk, B., 1997. Regional dependency of precipitation-altitude relationship in the Swiss Alps. Climatic Change 36 (3–4), 355–369.

Shu, C., Burn, D.H., 2004. Artificial neural network ensembles and their application in pooled flood frequency analysis. Water Resources Research 40, W09301.

Shu, C., Ouarda, T.B.M.J., 2007. Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space. Water Resources Research 43, W07438.

Siripitayananon, P., Chen, H.-C., Jin, K.-R., 2003. A modified nearest neighbors approach for estimating missing wind data. International Journal of Smart Engineering System Design 5 (3), 149–160.

Solomatine, D.P., Xue, Y., 2004. M5 model trees and neural networks: application to flood forecasting in the upper reach of the Huai River in China. Journal of Hydrologic Engineering 9, 491–501.

Teegavarapua, R.S.V., Chandramouli, V., 2005. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. Journal of Hydrology 312, 191–206.

Thornton, P.E., Running, S.W., White, M.A., 1997. Generating surfaces of daily meteorological variables over large regions of complex terrain. Journal of Hydrology 190 (3–4), 214–251.

USEPA, 2001. Better Assessment Science Integrating Point and Nonpoint Sources (BASINS) Version 3.0, Washington, DC.

Willmott, C.J., Robeson, S.M., Feddema, J.J., 1994. Estimating continental and terrestrial precipitation averages from rain-gauge networks. International Journal of Climatology 14 (4), 403–414.

Wu, C.L., Chau, K.W., Li, Y.S., 2008. River flow prediction based on a distributed support vector regression. Journal of Hydrology 358, 96–111.

Xia, Y.L., Fabian, P., Stohl, A., Winterhalter, M., 1999. Forest climatology: estimation of missing values for Bavaria, Germany. Agricultural and Forest Meteorology 96 (1–3), 131–144.

Young, K.C., 1992. A 3-way model for interpolating for monthly precipitation values. Monthly Weather Review 120 (11), 2561–2569.