Scientia Iranica A (2016) 23(1), 45–53

Sharif University of Technology
Scientia Iranica
Transactions A: Civil Engineering
www.scientiairanica.com

Three-dimensional imputation of missing monthly river flow data

F. Dikbaş*

Department of Civil Engineering, Pamukkale University, Denizli, P.O. Box 20020, Turkey.

Received 20 January 2014; received in revised form 7 April 2015; accepted 30 June 2015

KEYWORDS
River flow estimation;
3D imputation;
Büyük Menderes River;
Missing data;
Data-driven modeling.

Abstract. Imputation of missing data is a critical part of accurate data analysis and modeling. This paper presents 3D imputation as a new data-driven methodology to estimate missing values in time series data. The method is based on the assumption that all the observed data in a time series are related with each other and with data of the some other series. The available data is placed in a three-dimensional space so that the increasing or decreasing relationships between the observed data are appropriately represented. For the estimation of each missing value, the method searches and determines the best possible group of estimator data within the data space. Different data groups are found and used for the estimations of each individual group of missing data. The method is validated by removing and estimating all the observed monthly flow data of Sarayköy station on Büyük Menderes River in Turkey. Data of the downstream Burhaniye station constituted the second data layer in the model. High correlation values were obtained for all years between observations and estimations and the missing data of Sarayköy station was also estimated by using the proposed method.

1. Introduction

"Everything in the environment is connected to everything else" says the first informal law of ecology [1]. This interconnectedness and complexity causes difficulty in the estimation and modeling of an environmental variable. Most of the times, the researchers also experience the problem of incomplete or unobserved data as it is impossible to completely measure and record all variables in the environment. The requirement of data encouraged and forced researchers to develop numerous models for the estimation of incomplete or unobserved data. Physically based models and data-driven models constitute the two approaches used in river flow modeling.

The physically based models require sufficient understanding, assessment, and modeling of the properties and processes influencing and generating the flow. The accuracy of the results largely depends on the detail level of the descriptions of the influential parameters which are highly variable within the catchment. The requirement of increased accuracy causes increased complexity and demands collection of huge amounts of various concurrent data ranging in time and space. Inaccurate and unreliable results might be obtained from the conventional physically-based models when these requirements are not met [2].

Data-driven approaches make use of the information contained in the observed data mostly without considering the physical processes. The applicability of data-driven methods substantially increased with the increase of available observed data and the rapid developments in computational power. Of these methods, autoregressive integrated moving average (ARIMA) time series model [3-6] and Artificial Neural Networks (ANN) [7-13] are widely used for river flow estimations. Support Vector Machines (SVM), proposed by

*. *Tel.: +90 258 2963403; Fax: +90 258 2963460*
E-mail address: f_dikbas@pau.edu.tr

Vapnik [14], is also successfully applied in hydrological studies [15-18].

In most of the existing studies, river flows are regarded as one-dimensional series composed of single observations. Estimation of flow at one location by using the upstream records is rare in literature [19]. To the best of the author's knowledge, there is no study linking the observed values in a station with the data of neighboring stations in a three dimensional data space.

This paper describes a new data-driven approach to estimate the missing values in time series. The main idea behind the model is that each observation is mostly related with the nearest neighboring observations in the same station and the neighboring stations. To generate a three dimensional data space, a matrix is generated for each station and the matrix of each station covers a two-dimensional layer in the data space.

The method tries to find the best possible estimation for a missing value by choosing the best estimator data group in the data space. This is accomplished by deliberately removing and estimating the nearest neighbors of a missing value. All rectangular data groups in the space covering the missing node are evaluated to find the best estimations. The group giving the best results is then used for the estimation of the missing value. The process is repeated for estimating the remaining missing values. Apart from the existing methods mostly using the same estimator data group for all estimations, the method determines different estimator data groups for each missing value. The method evaluates all existing data without any smoothing of extreme values or ignoring any observation. In the testing phase of the method, significant correlations were obtained between the observed and the estimated values. Then, the missing flow data of Sarayköy station on Büyük Menderes River were estimated by incorporating the data of the downstream Burhaniye station.

2. Methodology

2.1. 3D imputation method

It is well known that every river basin has a certain coherence in that various portions of the basin relate to each other in a reasonably consistent way [20]. Thus, all observations of different stations on the same river basin are interrelated with each other up to a degree. Based on this assumption, a new method, called 3D imputation, is developed to estimate missing values in a time series data. The method assumes that each observation in a time series data is related with every other observation in the same series and with the observations of neighboring stations.

In this section, details of the methodology are presented by explaining its application on the monthly mean flows of Sarayköy station (#07-81) on Büyük Menderes River (Figure 1). The observations cover the years from 1981 to 2000. All the values in 1993 and 2000 are missing (24/240; 10%). The observations of the downstream Burhaniye station (#712) have been included in the process as the second data layer and it has no missing values.

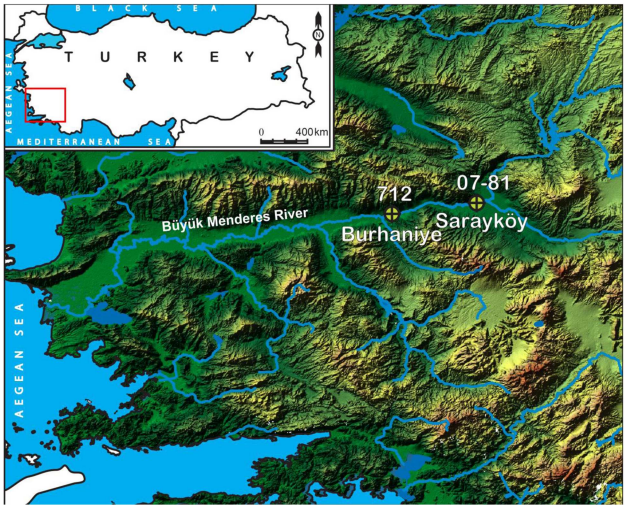

Figure 1. Locations of the observation stations.

The flowchart in Figure 2 shows the main steps of the proposed method. The software code of the method was written by making use of the interoperability feature of Microsoft Visual Basic and Microsoft Excel. This feature enables practical data acquisition and analysis and allows for reading data from running Excel spreadsheets, making computations and writing results back. Visual Basic was preferred to VBA to have the ability of compiling the program into an executable file.

The term 3D represents the three-dimensional data space generated by layers of data before the imputation process. For monthly river flow data, a layer consists of a station's observations located on a matrix with months in columns and years in rows (Figure 3). Each matrix is extended by locating the values of the last five months of the previous water year to the left (the blue region) and the next six months of the following water year to the right (the brown region). The years on the left are valid for the gray region. This enlargement is not compulsory, but it significantly simplifies software coding of the methodology.

2.2. The formulation

After the generation of the data space, the estimation process starts. First, the number of observed nearest neighbors for each cell is determined. Each cell has up to 8 nearest neighbors. The missing node with the highest number of nearest neighbors is estimated first. For the current sample, the missing node in October 1993 has 6 nearest neighbors as shown in Figure 3.

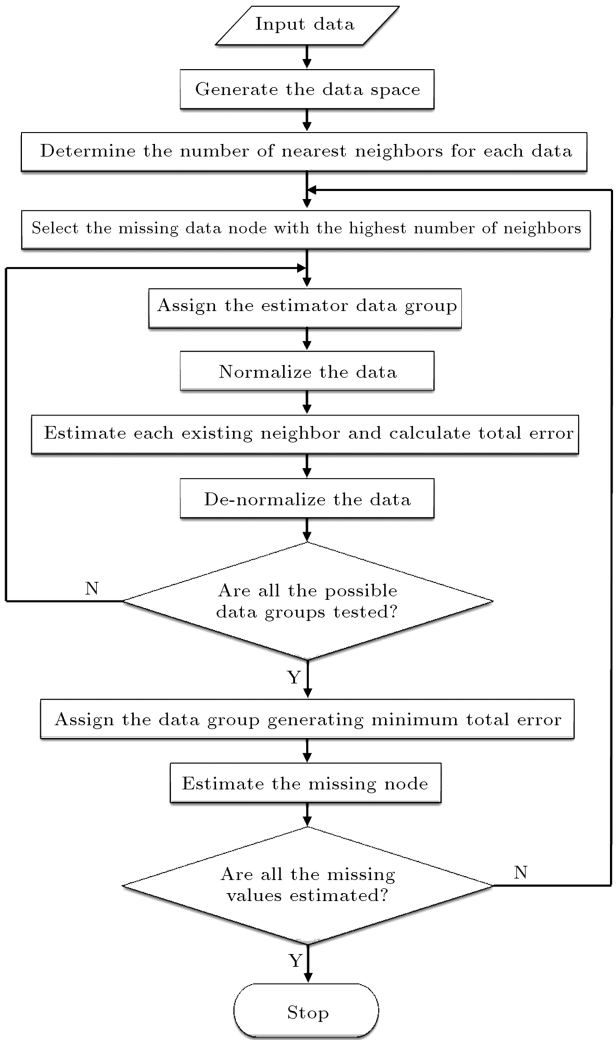

Figure 2. Flowchart of the 3D imputation method.

From this point, this node will be called the missing node.

The solution range for each cell covers five months before and after the cell in all years in all stations. The fundamental idea of the method is that the data group giving the best estimates for the nearest neighbors of a missing node will give the best estimation for the missing node itself. The search for the best estimator group is done within the solution range shown with a green frame in Figure 3. All possible rectangular data groups covering the missing node and its nearest neighbors are tested. The red frame in Figure 4 is the first data group to be tested for estimating the nearest neighbors of the missing node. The process of nearest neighbor estimation is repeated for all possible estimator groups. The video of the selection of estimator groups and the estimation of the first two missing values is provided at: http://youtu.be/wPZhfAWUPwo. Estimation of the first two missing values normally takes about 10 seconds on an average computer and the video is slowed down for increasing the understandability of the selection process.

Eq. (1) is used for all the estimations in the method. The equation is inspired by the first law of ecology stating that everything in the environment is connected to everything else. Consequently, if the flow values in a period are lower (or higher) than the anticipated values, then the missing values in that period will most probably be lower (or higher) than the other periods. The nearest observations have the highest impact on an observation and the relationship between the observations decreases with increase in the difference between observation times and locations.

$$Q = \frac{\sum_{i=1}^{n} \frac{Q_i^* w_i}{D_i^2}}{\sum_{i=1}^{n} \frac{w_i}{D_i^2}}. \tag{1}$$

In this equation, Q is the value to be estimated; n is the total number of existing observations in the assigned estimator data group in all stations; Q_i is the ith observed value in the estimator data group; w_i is the weight of the ith data; and D_i is the spatial distance of the ith data in the data space from the missing node.

For the current sample, the n value for the assigned estimator data group (shown with the red frame) is 179 (86 values from the Sarayköy station and 93 values from the Burhaniye station).

The weight value assigned to a cell is a measure of the contribution degree of the cell in the estimation of the missing value. The weight value for each cell is determined according to the neighborhood degree of a cell and the maximum neighborhood degree in the assigned data group. Consequently, each cell will have different weight values according to the position of the missing data and the width (or height) of the assigned data group. Figure 4(a) shows the neighborhood degrees of each cell in the solution range of the missing value. While the nearest neighbors are the 1st degree neighbors, the most distant neighbors are the 12th degree neighbors. The maximum neighborhood degree in the assigned data group is 12, which is the neighborhood degree of October and November 1981.

Table 1 shows the weight values to be assigned to each cell. These values are calculated according to Eq. (2):

$$w_i = \frac{100}{n_d^* d_{\max}}. \tag{2}$$

The numbers in the left column (n_d) of Table 1 are the highest possible numbers of cells that can be present in the 1st to the 14th neighborhood degrees. (There can be 8 1st degree neighbors and 112 14th degree neighbors.) The number of neighborhood degrees (d) is not limited to 14 and can increase according to the available number of observation years. The numbers

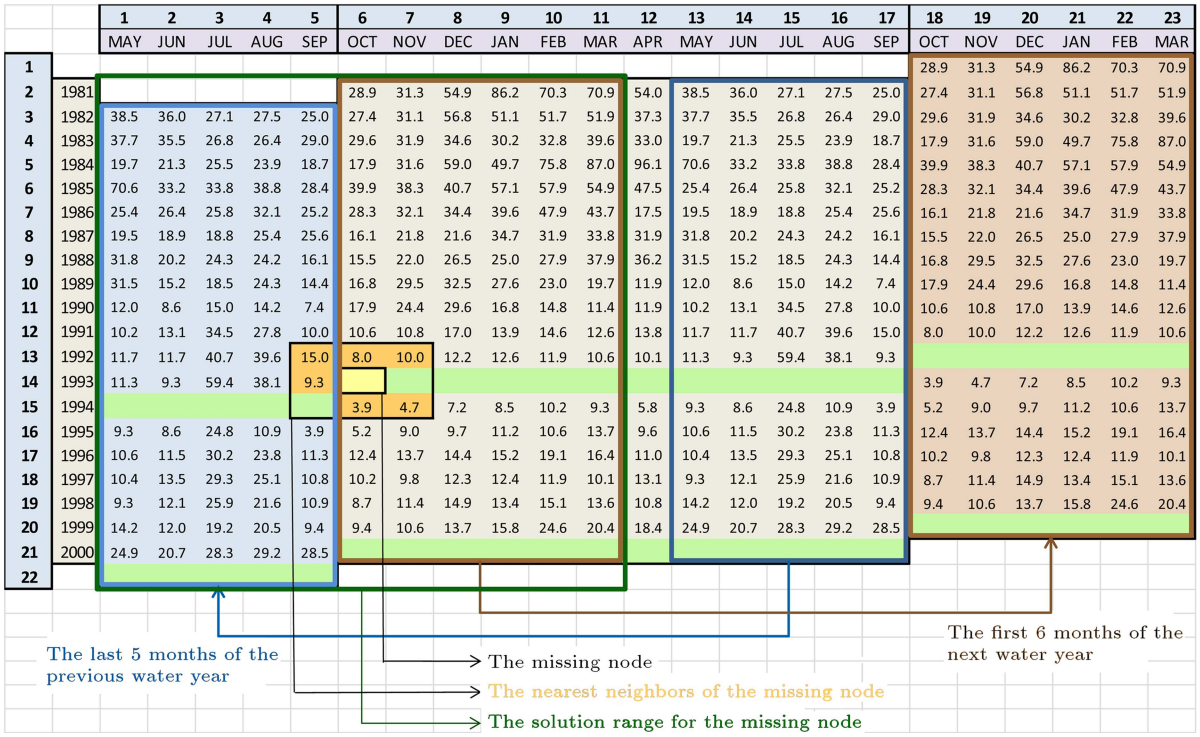

Figure 3. The extended data matrix; data from the Sarayköy station.

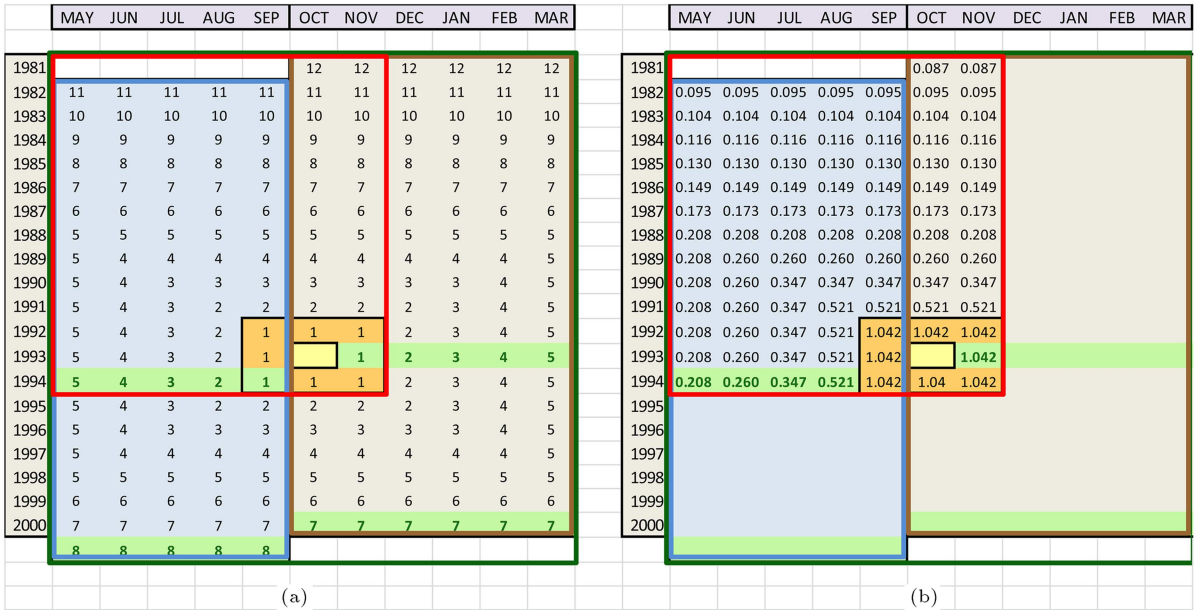

Figure 4. The neighborhood structure of a missing node (a) and the weights of its neighbors (b).

in the top row of Table 1 are the highest neighborhood degree values ($d_{\max}$) in the assigned data groups. (This value is 12 for the current example as stated above.) The total value of the weight distributed among the cells of an assigned range can be 100 if all cells of all neighborhood degrees are present in the assigned group. For each column, the total value of the weights multiplied with the number of elements in the first column equals to 100. For the current sample, the weights of the cells are taken from the 12th column of Table 1. The weights of the 1st degree neighbors are therefore 1.042, and 12th degree neighbors take the weight of 0.087 (Figure 4(b)).

2.3. The estimation process

First, the nearest neighbors of the missing node are removed from the set one by one and the data group giving the best estimations for them is determined.

Table 1. The weights distributed to each neighbor according to the degrees of neighborhood.

Number of elements in the degree (n_d)	Number of maximum neighborhood degree (d_{max})													
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
8	12.500	6.250	4.167	3.125	2.500	2.083	1.786	1.563	1.389	1.250	1.136	1.042	0.962	0.893
16		3.125	2.083	1.563	1.250	1.042	0.893	0.781	0.694	0.625	0.568	0.521	0.481	0.446
24			1.389	1.042	0.833	0.694	0.595	0.521	0.463	0.417	0.379	0.347	0.321	0.298
32				0.781	0.625	0.521	0.446	0.391	0.347	0.313	0.284	0.260	0.240	0.223
40					0.500	0.417	0.357	0.313	0.278	0.250	0.227	0.208	0.192	0.179
48						0.347	0.298	0.260	0.231	0.208	0.189	0.174	0.160	0.149
56							0.255	0.223	0.198	0.179	0.162	0.149	0.137	0.128
64								0.195	0.174	0.156	0.142	0.130	0.120	0.112
72									0.154	0.139	0.126	0.116	0.107	0.099
80										0.125	0.114	0.104	0.096	0.089
88											0.103	0.095	0.087	0.081
96												0.087	0.080	0.074
104													0.074	0.069
112														0.064

All possible rectangular data groups within the green solution range are tested for estimating the neighbors of the missing data. The groups which do not include the missing node should not be included in the evaluation.

The maximum horizontal range of an estimator data group includes 5 columns before and after the missing value and the maximum vertical range covers all the rows of the data matrix (all the years in the set are taken into account). The smallest estimator group is the first degree neighbors of the missing node. For the current data set, the number of possible estimator data groups varies between 500 and 2750 according to the location of the missing node in the data matrix. 2600 different estimator data groups were evaluated for estimation of the first degree neighbors of the missing node in October 1993.

The red frame in Figure 4 shows the first assigned estimator group among the tested 2600 data groups. The neighborhood degrees and weights of each cell in the estimator group are assigned as in Figure 4. For each missing node, the neighborhood degrees and the values of weights in the estimator groups vary according to the location of the missing node and the size of the estimator group.

Generally, a downstream increase is observed in the river flow rates. To eliminate the negative impacts of the rate differences between the stations, a normalization procedure is applied. Each layer in the estimator data group should be normalized separately. First, the maximum value in each layer of the selected estimator data group is determined. (The values outside the range of the estimator group should not be considered.) Then, each layer is separately normalized by using Eq. (3):

$$Q_{new} = \frac{Q}{Q_{max}}. \tag{3}$$

In the above equation, Q is the data to be normalized, Q_{max} is the maximum value in the estimator group in the current layer, and Q_{new} is the normalized value between 0 and 1. After the application of Eq. (3) on all elements of the estimator group, the maximum value in each layer becomes 1.

Eq. (1) is used for estimating the nearest neighbors of the missing node and all the observed values of all stations within the estimator group are included in the calculation. The estimation success of each estimator group is assessed by determining the sum of the absolute differences between the estimated and observed values. The group giving the least total difference is selected as the best estimator group.

The data group giving the best estimates for the neighbors of October 1993 covers the range between July 1982 and March 1994 on the extended data matrix. This best estimator group is then used in Eq. (1) to estimate the missing node in October 1993.

After estimation of the first node, the estimation process is repeated for each group of remaining missing data. In the current example, the solution range moves to the right together with the missing node until all missing data in 1993 is estimated. Then, the flow rates in 2000 are estimated.

As each estimator group might have different maximum values, the normalized data should be de-normalized before the assignment of a new estimator group. De-normalization turns the normalized values back to original observed values. The normalization and de-normalization steps are not required if there is one observation series.

3. Results and discussion

3.1. Evaluation of the model performance

The accuracy and reliability of the proposed method are tested by making estimations for all of the observed flow rates in Sarayköy station. The observed data for each year is deliberately removed and estimated separately. The estimated values are compared with the observed data. This yearly comparison enables the assessment of estimation success of the model throughout the series instead of evaluating the series with a single statistic for the whole series.

The 3D imputation method was successful in estimating the observed values throughout the whole period, though the flow rates of the river show a downward trend (Figure 5). The statistical evaluation of the obtained results is made by calculating the coefficient of correlation (R), Nash-Sutcliffe efficiency coefficient (E), normalized root mean square error (NMRSE), and Mean Absolute Percentage Error (MAPE) for each modeled year and for the whole series (Table 2). The estimations of the missing values in 1993 and 2000 are not included in the statistical evaluation of the estimation performance.

The obtained correlation coefficient for the whole series is 0.973, while the maximum annual correlation is 0.998 in 1991 and 1995, and the minimum is 0.846 in 1998. Of all the 18 annual correlations, 6 (33%) were over 0.99, 12 (67%) were over 0.95, and 16 (89%) were over 0.91.

The Nash-Sutcliffe efficiency coefficient (E) is a normalized statistic that shows how well the plot of observed versus estimated data fits the $y = x$ line. $E = 0$ indicates that the estimations are as accurate as the mean of the observed data and if $E < 0$, then the observed mean is a better predictor than the model. If $E = 1$, then the estimated data perfectly matches the observed data. The Nash-Sutcliffe efficiency coefficient obtained for the estimations of the developed model in this study was 0.938. In the yearly evaluations, the maximum E value was obtained for 1991 (0.979) and the minimum was for 1998 (0.135). All E values were over 0 and 6 (33%) out of the 18 annual E values were over 0.9, while 13 (72%) of them were over 0.8.

Normalized Root Mean Square Error (NRMSE) is a normalized measure of the average magnitude of the estimation errors. It ranges from 0 to infinity, with 0 being a perfect score. The data in this study has higher flow values in the earlier years and the decreasing trend causes scale differences between years. Thus, NRMSE is preferred to RMSE which would give deceptive values for the data set in this study. The minimum (0.035) and the maximum (0.155) NRMSE values were obtained for the data of 1981 and 1998, respectively. NRMSE for the whole series was 0.042. Of the annual NRMSE values, 15 (83%) were under 0.1.

The Mean Absolute Percentage Error (MAPE) is an unbiased measure of the accuracy of a method for estimating fitted time series data. The lowest possible

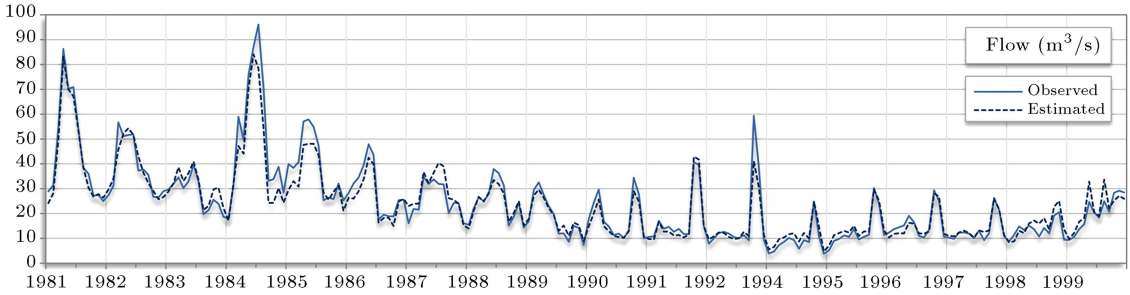

Figure 5. The observed and estimated flow rates of Sarayköy station (07-81).

Table 2. Annual and whole estimation performance metrics for the 3D imputation method.

	1981	**1982**	**1983**	**1984**	**1985**	**1986**	**1987**	**1988**	**1989**	**1990**
R	0.993	0.926	0.949	0.977	0.941	0.981	0.916	0.977	0.995	0.994
E	0.977	0.858	0.747	0.853	0.705	0.858	0.547	0.915	0.949	0.882
NRMSE	**0.035**	0.072	0.079	0.097	0.114	0.078	0.127	0.058	0.055	0.081
MAPE	**5.8**	7.7	10.0	15.0	13.7	9.7	14.1	6.6	11.0	10.3
	1991	**1992**	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**	**Whole**	
R	**0.998**	0.995	0.983	**0.998**	0.957	0.986	**0.846**	0.867	**0.973**	
E	**0.979**	0.847	0.826	0.933	0.891	0.951	**0.135**	0.658	**0.938**	
NRMSE	0.036	0.099	0.090	0.057	0.063	0.042	0.155	0.133	**0.042**	
MAPE	7.9	13.0	**27.8**	16.4	11.3	6.6	21.2	13.3	**12.3**	

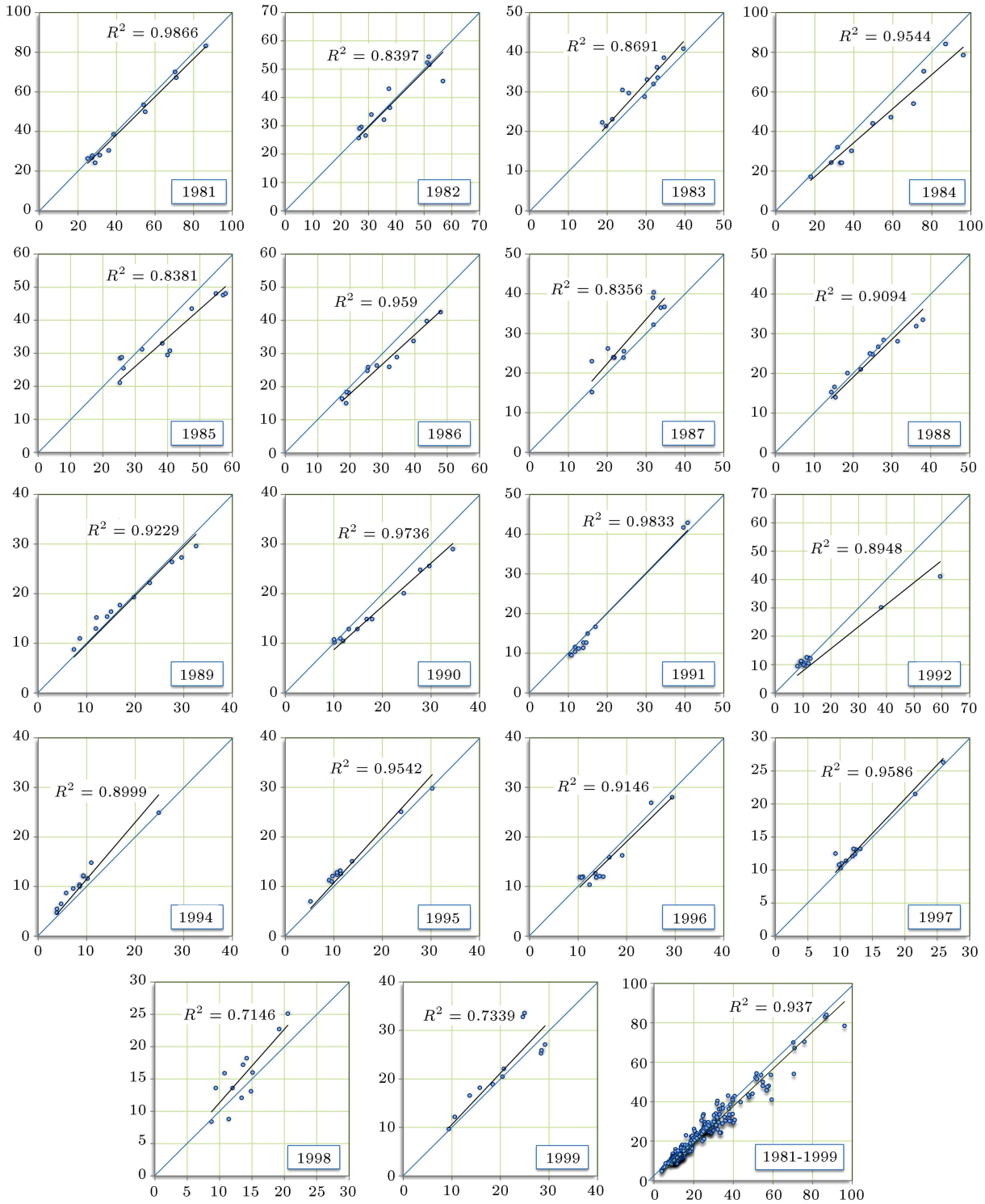

Figure 6. The scatter graphs of observed (horizontal axis) versus estimated (vertical axis) flow rates of Sarayköy station.

value of MAPE is 0, indicating a perfect fit, and it has no upper limit. The lowest annual MAPE value was in 1981 (5.8%), while the highest was in 1994 (27.8%). Throughout the whole series, the lowest 17 (7.9%) monthly MAPEs were below 1% and the lowest 57 (26.4%) were below 5%. The calculated MAPE values for the 133 months (62%) out of the existing 216 months were below the MAPE value (12.3%) calculated for the whole series.

It appears that the method can produce reliable results, but it is noticed that though satisfactory, three of the four measures have their worst values for the estimations of observed values in 1998 (Figure 6). Investigation of the data reveals that the cause for this

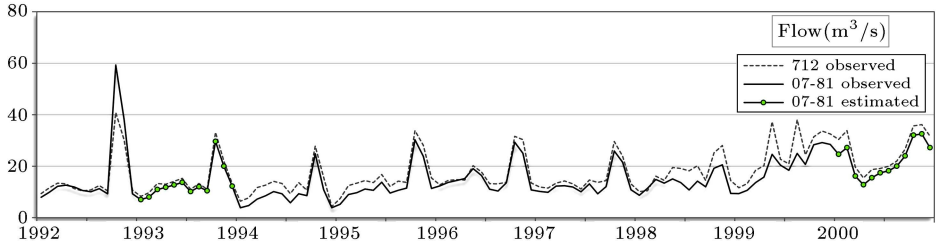

Figure 7. Observed and estimated values for stations 712 and 07-81.

situation might be that the lowest observed flow values are in 1998.

It is believed that the estimation power of the method might be increased by making developments on the following subjects:

- Non-rectangular estimator data group shapes might be used to increase the number of tested estimator data groups;
- The criteria for the selection of the best estimator group might be modified to decrease the difference between the observed and estimated values;
- The logic for distribution of weights to the neighboring data might be improved;
- This paper only makes use of the data of two stations, but in further studies, a higher number of stations and other related variables might be included in the research.

3.2. Estimation of missing data

The 24 missing values of the years 1993 and 2000 were estimated by using the 3D Imputation method. Figure 7 shows the observed data of both stations between 1992 and 2000 together with the estimated data. Naturally, the flow rates of the downstream Burhaniye station (712) are higher than the flows of Sarayköy station (07-81). The imputed values are shown with circles and have a good fit to the general flow pattern.

The obtained results show that the proposed 3D imputation model accurately estimates both the observed and the missing river flow records. As mentioned in the Introduction, so far no one appears to have applied the approach of 3D imputation method to the field of missing value analysis. It is believed that the method might also be applied in a wide range of disciplines other than hydrology.

4. Conclusions

This study provides the framework of 3D imputation method, proposing it as a new way to estimate missing values in time series data. The essential advantages of the method are:

- It has the ability to select the best estimator group among all existing data without ignoring any values, including the extremes;
- A simple formulation is used for both the selection of the best estimator group and the estimation of missing data;
- No personal intervention or decision making is required during the whole process;
- A short computation time is required;
- The method is successful in estimating the observed river flow data of Sarayköy station.

The high correlations obtained between the observed and estimated river flow data are very promising and future work will concentrate on the development of additional features for the software, such as inclusion of a higher number of estimator groups, assigning variable weights to the neighbors, and other possibilities for improving the performance of the model. Some areas of future work will be to apply the model on hydrologic variables like precipitation, temperature, sediment transport, etc. and on other disciplines like economy and medicine (especially public health and biostatistics). We believe that the proposed 3D imputation method may improve knowledge about missing data analysis.

Acknowledgements

We would like to thank The General Directorate of State Hydraulic Works and The General Directorate of Renewable Energy in Turkey for providing the data used in this study.

References

1. Commoner, B., *The Closing Circle: Nature, Man, and Technology*, Knopf, ISBN 978-0-394-42350-0, New York (1971).
2. Firat, M. and Turan, M.E. "Monthly river flow forecasting by an adaptive neuro-fuzzy inference system", *Water and Environment Journal*, **24**, pp. 116-125 (2010).
3. Abrahart, R.J. and See, L. "Comparing neural network (NN) and autoregressive moving average (ARMA) techniques for the provision of continuous river flow

forecasts in two contrasting catchments", *Hydrological Processes*, **14**, pp. 2157-2172 (2000).

4. See, L. and Openshaw, S. "Using soft computing techniques to enhance flood forecasting on the River Ouse", *Proceeding Hydroinformatics'98: 3rd International Conference on Hydroinformatics*, Copenhagen, Denmark, **24-26**, pp. 819-824 (August 1998).
5. Yurekli, K., Kurunc, A. and Simsek, H. "Prediction of daily streamflow based on stochastic approaches", *Journal of Spatial Hydrology*, **4**(2), pp. 1-12 (2004).
6. Modarres, R. "Streamflow drought time series forecasting", *Stochastic Environmental Research and Risk Analysis*, **21**, pp. 223-233 (2007).
7. Dolling, O.R. and Varas, E.A. "Artificial neural networks for streamflow prediction", *Journal of Hydraulic Research*, **40**(5), pp. 547-554 (2002).
8. Kisi, O. "River flow forecasting and estimation using different artificial neural network technique", *Hydrology Research*, **39**(1), pp. 27-40 (2008).
9. Wang, W.C., Chau, K.W., Cheng, C.T. and Qui L. "A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series", *Journal of Hydrology*, **374**, pp. 294-306 (2009).
10. Kentel, E. "Estimation of river flow by artificial neural networks and identification of input vectors susceptible to producing unreliable flow estimates", *Journal of Hydrology*, **375**, pp. 481-488 (2009).
11. Keskin, M.E. and Taylan, D. "Artificial models for interbasin flow prediction in southern Turkey", *Journal of Hydrologic Engineering*, **14**(7), pp. 752-758 (2009).
12. Huo, Z., Feng, S., Kang, S., Huang, G., Wang, W. and Guo, P. "Integrated neural networks for monthly river flow estimation in arid inland basin of Northwest China", *Journal of Hydrology*, **420-421**, pp. 159-170 (2012).
13. Okkan, U. "Wavelet neural network model for reservoir inflow prediction", *Scientia Iranica A*, **19**(6), pp. 1445-1455 (2012).
14. Vapnik, V., *The Nature of Statistical Learning Theory*, Springer, New York (1995).
15. Asefa, T., Kemblowski, M., McKee, M. and Khalil, A. "Multitime scale stream flow prediction: The support vector machines approach", *Journal of Hydrology*, **318**, pp. 7-16 (2006).
16. Kalteh, A.M. and Hjorth, P. "Imputation of missing values in a precipitation-runoff process database", *Hydrology Research*, **40**, pp. 420-432 (2009).
17. Lin, J.Y., Cheng, C.T. and Chau, K.W. "Using support vector machines for long-term discharge prediction", *Hydrological Sciences Journal*, **51**(4), pp. 599-612 (2006).
18. Samsudin, R., Saad, P. and Shabri, A. "River flow time series using least squares support vector machines", *Hydrology and Earth System Sciences*, **15**, pp. 1835-1852 (2011).
19. Turan, M.E. and Yurdusev, M.A. "River flow estimation from upstream flow records by artificial intelligence methods", *Journal of Hydrology*, **369**, pp. 71-77 (2009).
20. Leopold, L.B. and Skibitzke, H.E. "Observations on unmeasured rivers", *Geografiska Annaler*, **49A**(2-4), pp. 247-255 (1967).

Biography

Fatih Dikbaş received his MS degree in Civil Engineering from Istanbul Technical University in 1994 and PhD degree from Pamukkale University in 2002. Currently, he is an Associate Professor in the Civil Engineering Department at Pamukkale University. His main research areas are hydrology, data-driven modeling, statistical analysis, hydromechanics, and software development.