

PAPER • OPEN ACCESS

An improved hybrid data-driven model and its application in daily rainfall-runoff simulation

To cite this article: Guangyuan Kan *et al* 2016 *IOP Conf. Ser.: Earth Environ. Sci.* **46** 012029

View the [article online](#) for updates and enhancements.

You may also like

- [Recent progress on the poly\(arylene ether\)s-based electrospun nanofibers for high-performance applications](#)
Pan Wang, Xidi Liu, Dengyu Wang et al.
- [Optimization of high performance hybrid composite under ballistic test and performance of the hybrid composite under IED blast](#)
Sravan Kumar Khuntia, Kumbhar Avinash Suresh, Jithin Joshi et al.
- [Electro-optical Properties and Temporal Stability of the Guest-Host DCNP/PEK-c Polymer Thin Film](#)
Wei Shi, Chang-Shui Fang, Qi-Wei Pan et al.

An improved hybrid data-driven model and its application in daily rainfall-runoff simulation

Guangyuan Kan^{1,2}, Xiaoyan He¹, Liuqian Ding¹, Jiren Li¹, Tianjie Lei¹, Ke Liang³, Yang Hong^{2,4}

¹State Key Laboratory of Simulation and Regulation of Water Cycle in River Basin, Research Center on Flood & Drought Disaster Reduction of the Ministry of Water Resources, China Institute of Water Resources and Hydropower Research, Beijing 100038, China

²State Key Laboratory of Hydrosience and Engineering, Department of Hydraulic Engineering, Tsinghua University, Beijing, China

³College of Hydrology and Water Resources, Hohai University, Nanjing 210098, China

⁴Department of Civil Engineering and Environmental Science, University of Oklahoma, Norman, OK, United States

E-mail: kanguangyuan@126.com

Abstract. In previous literatures, a coupled data-driven rainfall-runoff (RR) model, NU-PEK, has been proposed and successfully applied in hourly RR simulation task. However, numerical experiments show that its performance for daily RR simulation is unsatisfactory. It is noticed that the poor performance is due to the inability of the original model to capture the much higher non-linear characteristics contained in the daily data. In order to improve the non-linearity simulation capability of the original model, an improved model named NU-PKEK and its calibration methodology are developed in this paper. The improved model is constituted by adding a K-means clustering module and utilizing multiple NU-PEK modules instead of using only one NU-PEK model. This study applies the improved model, the Xinanjiang model, and the original model for daily RR simulation in Chengcun catchment for intercomparison and verification. The simulation results prove that the NU-PKEK performs best, and has better simulation and forecasting capability.

1. Introduction

Data-driven hydrological models have been proved to be effective, simple, and concise methods in recent years (Piotrowski et al., 2013). They have been used widely in hydrological forecasting, water resources management, and environmental simulation (Shouke et al., 2012; Mohammad et al., 2013; Vahid et al., 2013). Applications indicate that the data-driven model is an effective and robust method for the RR simulation and forecasting on the up-dating mode, which can forecast discharge one-time-step ahead. On the other hand, multi-time-step simulation is also studied by some hydrologists (Li et al., 2014; Kan et al., 2015).

The recently proposed NU-PEK model (Kan et al., 2015) aims at the implementation of non-updating RR simulation with higher accuracy and better stability than traditional conceptual models



such as the Xinanjiang and the IHACRES models. Nevertheless, numerical experiments show that when applying the NU-PEK model for the daily RR simulation tasks, the model performance is degraded. For the purpose of ameliorating the non-linearity simulation capability of the NU-PEK model and improving model performance in daily RR simulation, an improved model named NU-PKEK and its calibration methodology are developed in this paper. The NU-PKEK model is constituted by adding a K-means clustering module into the original NU-PEK model and utilizing multiple NU-PEK modules instead of using only one NU-PEK model. The input vectors are classified into several categories by the K-means clustering module to weaken the non-linearity contained in the input data. Multiple NU-PEK modules are adopted to simulate their corresponding category of input-output data to generate more precisely forecasted results. This study applies the improved model, the Xinanjiang model, and the original model for daily RR simulation in Chengcun catchment for the purpose of comparison and verification.

2. Study area and data description

We carry on the daily rainfall-runoff simulation in Chengcun catchment (290km²), which lies in the Qiantang River basin, Anhui province, China. There are ten rainfall gauges located in this area. Observed daily rainfall, evaporation, and average discharges range from 1986 to 1994 were utilized as the calibration data, while data from 1995 to 1999 were utilized as the validation data. The Thiessen polygon map for the Chengcun catchment are shown in Figure 1.



Figure 1. Thiessen polygon map for the Chengcun catchment.

3. Methodology

3.1. PKEK approximator

3.1.1. K-means clustering. The K-means clustering (Grigorios and Aristidis, 2014; Kapageridis, 2015; Rehab and Mohammed, 2015) is a famous and widely used partitioning and clustering method (MATLAB, 2012). The K-means method is usually calibrated or trained by the iterative method.

3.1.2. PEK approximator. The PEK approximator is a hybrid data-driven model and it can be applied for the simulation of the multi-input single-output (MISO) system mapping relationship. Detailed principle of the PEK approximator can be found in the corresponding literatures (Kan et al., 2015).

3.1.3. PKEK approximator. PKEK approximator is an improved version of the previously proposed PEK approximator. It is combined by the PMI (partial mutual information) based IVS (input variable selection) module, the K-means clustering based input variable clustering module, and the ENN (ensemble artificial neural network) and KNN (K-nearest neighbour) simulation module. The PKEK approximator can be seen as a hybrid approximator that composed by a K-means clustering module and multiple PEK modules. The architecture of the PKEK approximator is demonstrated in Figure 2.

The simulation and calibration method of the PKEK approximator is as follows: The input variables are selected by the PMI-based separate IVS scheme to generate the selected input vectors.

After that, each selected input vector is fed into the K-means clustering algorithm to determine which category it belongs to. For each selected input vector, we choose a corresponding PEK module to calculate the output. In the PEK module, the output is estimated by the ENN, and the output error is estimated by the KNN regression. The final simulated output is the sum of the estimated output and output error. The calibration of the PKEK approximator is almost the same with the PEK approximator. The difference is that we need to calibrate the K-means clustering before calibrate PEK modules by an iterative method (Grigorios and Aristidis, 2014; Kapageridis, 2015; Rehab and Mohammed, 2015).

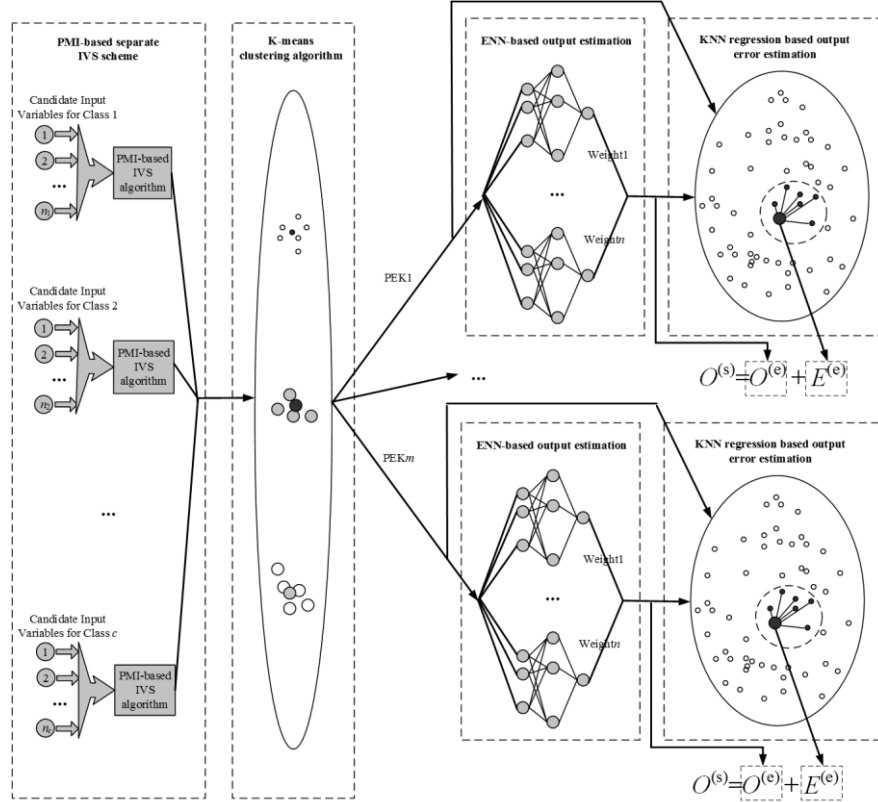


Figure 2. The architecture of the PKEK approximator.

3.2. NU-PKEK module

3.2.1. Non-updating RR simulation of the NU-PKEK model. The non-updating modelling approach of the NU-PKEK model is similar to the previously proposed NU-PEK model. The difference is the addition of the K-means clustering. The modelling approach of the NU-PKEK model is as follows:

$$X_t^{(Q_SIM)} = \left(Q_{t-1}^{(SIM)}, Q_{t-2}^{(SIM)}, \dots, Q_{t-n_Q}^{(SIM)} \right)^T \quad (1)$$

$$X_t^{(SWCR)} = \left(SWCRs_t^{(1)}, SWCRs_t^{(2)}, \dots, SWCRs_t^{(n_p)} \right)^T \quad (2)$$

$$X_t^{(S)} = \left(IVS_{Q_SIM} \left(X_t^{(Q_SIM)} \right), IVS_{SWCR} \left(X_t^{(SWCR)} \right) \right)^T \quad (3)$$

$$Q_t^{(EST)} = F_{EBPNN} \left[F_{K-means} \left(X_t^{(S)} \right) \right] \quad (4)$$

$$E_t^{(EST)} = F_{KNN} \left[Q_t^{(EST)}, F_{K-means} \left(X_t^{(S)} \right) \right] \quad (5)$$

$$Q_t^{(SIM)} = Q_t^{(EST)} + E_t^{(EST)} \quad (6)$$

where $X_t^{(Q_SIM)}$ denotes the candidate simulated antecedent discharge (SAD) input vector; $X_t^{(SWCR)}$ denotes the candidate sliding window cumulative rainfall (SWCR) input vector; $Q_{t-i}^{(SIM)}$ denotes the SAD;

$SWCRs_i^{(i)}$ denotes the SWCR with sliding window width i ; nQ and nP denote the order of the SAD and the SWCR respectively; IVSQ_SIM denotes PMI-based IVS for candidate SAD input vector; IVS_SWCR denotes PMI-based IVS for candidate SWCR input vector; $X_t^{(s)}$ denotes the selected input vector; $Q_t^{(EST)}$ denotes the estimated output discharge associated with $X_t^{(s)}$; $E_t^{(EST)}$ denotes the estimated output discharge error associated with $X_t^{(s)}$; $Q_t^{(SIM)}$ denotes the simulated output discharge and it is the final prediction of the output discharge; FEBPNN denotes the EBPNN discharge estimation; FKNN denotes the KNN discharge error estimation; FK-means denotes the K-means clustering. The non-updating simulation procedure of the NU-PKEK model for each flood event is almost the same with the NU-PEK model. The difference is that after the PMI-based separate IVS, the K-means clustering is executed to classify the input vector and fed it to different PEK module to calculate the forecasted discharge value.

3.2.2. Calibration of the NU-PKEK model. The calibration of the NU-PKEK model is almost the same with the NU-PEK model. The difference is the addition of the calibration of the K-means clustering. The K-means algorithm is calibrated by the iterative method and this calibration process is repeated for 100 times with different initial parameters to avoid the local minimum issue. The optimal number of classes is determined by the Silhouette value. In this paper, approximately 70% of the flood events are utilized to create the calibration data and the remaining events are utilized to create the validation data.

3.3. Xinanjiang model

The Xinanjiang model (Zhao et al., 1980, 1984, 1992) is a conceptual RR model which has been successfully and widely used in China. The flowchart of the Xinanjiang model is demonstrated in Figure 3 (Yao et al., 2009).

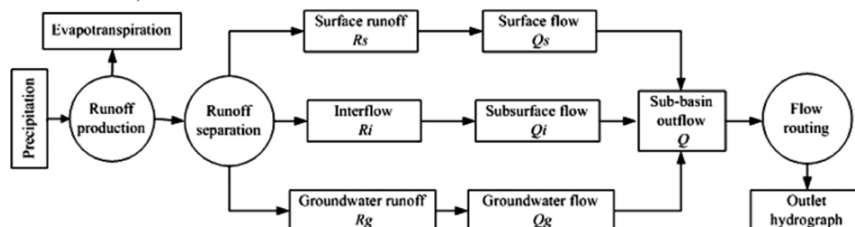


Figure 3. Xinanjiang model flow chart (Yao et al., 2009).

4. Results and discussion

4.1. Calibration of the NU-PKEK and NU-PEK models

4.1.1. Input variable selection. Considered that we are studying the daily simulation, we set the order of SWCR and SAD (i.e. nP and nQ) equal to 10, and 10 is large enough and rational for daily RR simulations. This means that the rainfall will becomes runoff with a lag at most 10 days. After the IVS we have found that the lag time of most selected SADs are less than 10. Therefore, this result conform that the order of 10 is large enough to describe the mapping relationship between the daily rainfall and runoff data of the Chengcun catchment.

4.1.2. K-means clustering module. We use the Silhouette value to determine the optimal cluster number. The Silhouette refers to a method of interpretation and validation of consistency with clusters of data. The technique provides a succinct graphical representation of how well each object lies within its cluster (Rousseeuw, 1987). The classification results of the K-means clustering algorithm is shown in Figure 4. Figure 4 (A) demonstrates the relationship between the cluster number and the Silhouette value. It can be seen from the figure, when we dividing the input samples into 3 clusters, the

Silhouette value reaches the maximum value. Therefore, we divide the input sample into 3 clusters. Figure 4 (B) shows the Silhouette value distribution when the input samples are divided into 3 clusters. We can see that the Silhouette value distributes evenly and the clustering results is rational good and acceptable.

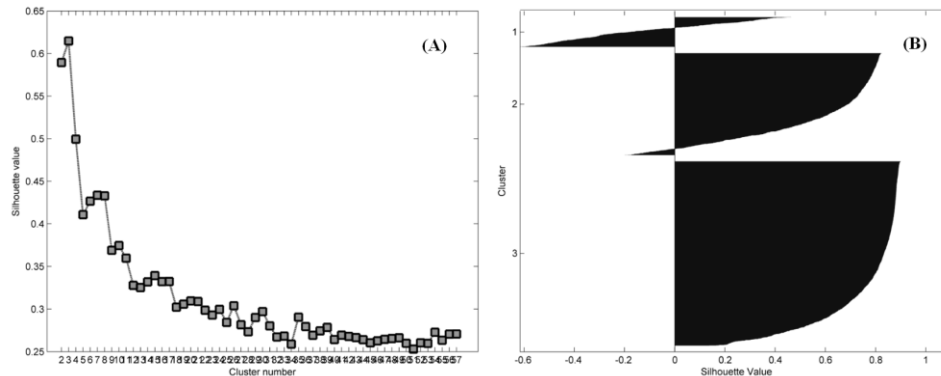


Figure 4. The classification results of the K-means clustering algorithm. (A) Relationship between the cluster number and the Silhouette value; (B) The Silhouette value distribution when the samples are divided into 3 clusters.

4.1.3. The PEK module. The PEK module is calibrated by a NSGA-II, Levenberg-Marquardt (LM), and leave-one-out cross-validation coupled method. The arithmetic parameters of the NSGA-II algorithm are set as: population size pop=90, evolution generation total number gen=1000, crossover probability pc=0.85, and mutation probability pm=0.15. The arithmetic parameters of the LM method are set as follows: minimum performance gradient min_grad=1-10, initial mu=0.001, mu decrease factor mu_dec=0.1, mu increase factor mu_inc=10, and maximum mu mu_max=110. The lower and upper boundaries of K for the KNN algorithm are set to 1 and 300.

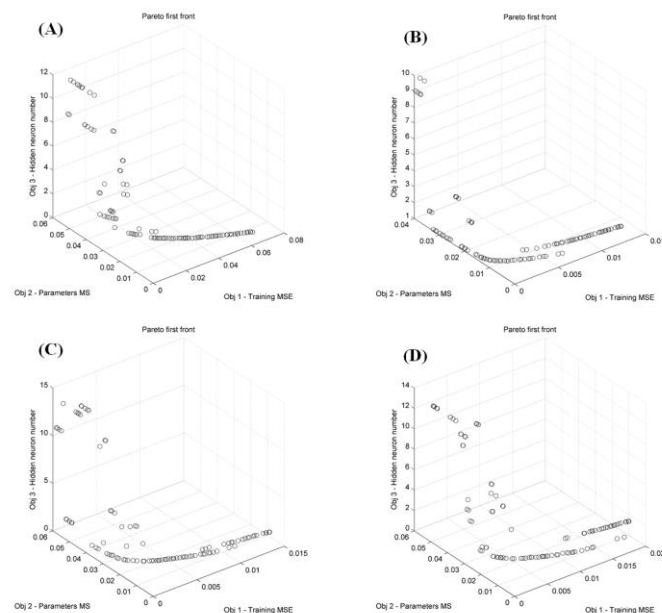


Figure 5. The optimized Pareto first front of the ENN. (A) The first cluster for the NU-PKEK model; (B) The second cluster for the NU-PKEK model; (C) The third cluster for the NU-PKEK model; (D) The NU-PEK model.

The optimized Pareto first front of the ENNs for the NU-PKEK and NU-PEK models are shown in Figure 5. Because we divide the input samples into 3 clusters, we construct three PEK modules for the NU-PKEK model. Figure 5 (A), (B), and (C) demonstrate the Pareto first front for the first, second,

and third cluster. Figure 5 (D) demonstrates the Pareto first front for the NU-PEK model. The objective function used for the NSGA-II algorithm is mean squared error (MSE), mean squared value of the network parameters (MS), and hidden layer neuron number. As we can see, all of the four Pareto fronts distribute evenly which means that the optimization results is reasonable. It can be seen that the configuration of the Pareto fronts in (D) seems like a combination of Pareto fronts in (A), (B), and (C). This fact implies that the improved model ((A) + (B) + (C)) contains the simulation capability of the original model (only (D)) and can simulate more precisely by utilizing three Pareto fronts.

4.2. Calibration of the Xinanjiang model

The Xinanjiang model parameters are optimized by the SCE-UA (Duan et al., 1992, 1993, 1994) method. The algorithmic parameters are set as: p (number of complexes)=15, maxn (maximum number of objective function evaluations)=100000, kstop (number of shuffling loops in which the objective function must improve by the specified percentage pcento)=10, pcento (percentage by which the objective function must change in the specified number of shuffling loops kstop)=0.1, and peps (minimum parameter space allowed)=0.001, respectively.

4.3. Model performance comparison

We use the scatter plots of the observed and simulated discharges and the regression R values to inspect the overall performance of the three models (see Figure 6).

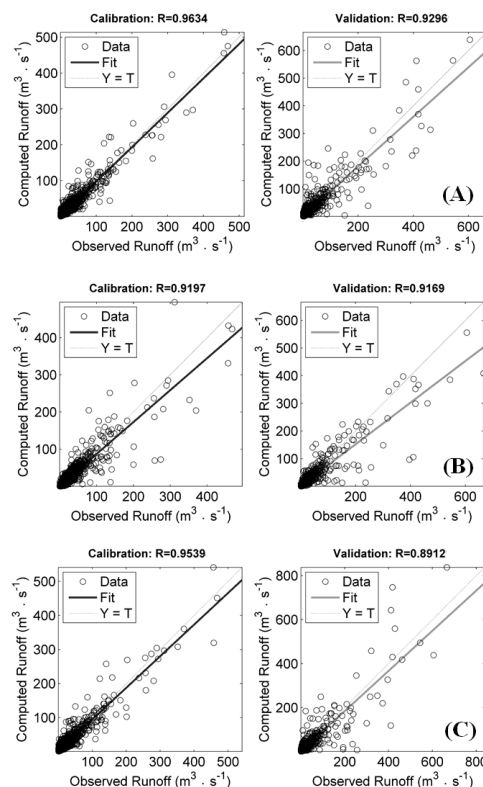


Figure 6. The scatter plots of the observed and simulated discharges and the regression R values: (A) NU-PKEK model, (B) Xinanjiang model, (C) NU-PEK model.

As demonstrated in Figure 6, as for the calibration period, the NU-PKEK model obtains the best result ($R = 0.9634$), the NU-PEK model obtains the moderate result ($R = 0.9539$), and the Xinanjiang model obtains the worst result ($R = 0.9197$). It can be noticed that the data scatters of the NU-PKEK model intensively lie close to the 45 degree line. The distribution of the data scatters of the NU-PKEK model is relatively more uniform for small, middle, and large discharge values. These results indicate that the NU-PKEK model generates the best result and obtains the most stable performance in the calibration period.

As for the validation period, the NU-PKEK model obtains the best result ($R = 0.9296$), the Xinanjiang model obtains the moderate result ($R = 0.9169$), and the NU-PEK model obtains the worst result ($R = 0.8912$). As demonstrated in the Figure 6, the data scatters of the NU-PKEK model distributed normally around the 45 degree line. This result indicates that the simulation results of the NU-PKEK model do not have bias to be larger or smaller and shows a very stable property. However, the simulated discharges of the Xinanjiang model is smaller than the observed values, while the simulated discharges of the NU-PEK model is larger than the observed values. This phenomenon becomes more obvious especially for large discharge values.

After analyzing the accuracy in calibration and validation period, the accuracy declination is also compared. The Xinanjiang model obtains the best accuracy declination value ($0.9169/0.9197 = 0.9970$), the NU-PKEK model obtains the moderate accuracy declination value ($0.9296/0.9634 = 0.9649$), and the NU-PEK model obtains the worst accuracy declination value ($0.8912/0.9539 = 0.9343$). The results show that although the NU-PKEK model obtains worse accuracy declination value than the Xinanjiang model, it outperforms the original NU-PEK model and generates better forecasted results.

The analysis of the three models shows that the NU-PKEK model can generate better overall results than the original NU-PEK model and the Xinanjiang model. Especially for peak values, the NU-PKEK model is much better than other two models. The comparison results indicate that the NU-PKEK model outperforms other models in Chengcun catchment and prove its satisfactory accuracy and stability.

5. Conclusions

An improved hybrid data-driven model, named NU-PKEK, is proposed in this paper to overcome the disadvantages of the previously proposed NU-PEK model in daily RR simulation. For the purpose of comparing performances of different models, the NU-PKEK, Xinanjiang, and NU-PEK models are applied to daily RR simulation in Chengcun catchment. It can be remarked that the NU-PKEK model is able to simulate generally much better runoff hydrograph than other models. From the application and analysis of the NU-PKEK model, we can conclude that:

(1) The NU-PKEK model have all the advantages of the previously proposed NU-PEK model, including better calibration and validation non-updating simulation accuracy, automatic calibration without much impact from human experiences, good compromise between the sufficiency and parsimony of the input data, and good compromise between the simulation accuracy and topology complexity, and so on.

(2) Compared with the original NU-PEK model, the performance of the NU-PKEK model becomes much better for the daily RR simulation. The improvement in the goodness-of-fit to the observed discharge of the NU-PKEK model is owed to the addition of the K-means clustering method combined with the utilization of multiple PEK modules. With the clustering method and multiple PEK modules, the NU-PKEK model can simulate the characteristics of the RR relationship much precisely and it's forecasting capability and stability becomes better.

References

- [1] Grigorios T., Aristidis L., 2014. The MinMax k-means clustering algorithm. *Pattern Recognition*. 47 (7), 2505-2516.
- [2] Kan G., Yao C., Li Q., Li Z., Yu Z., Liu Z., Ding L., He X., Liang K., 2015. Improving event-based rainfall-runoff simulation using an ensemble artificial neural network based hybrid data-driven model. *Stochastic Environmental Research and Risk Assessment*. DOI: 10.1007/s00477-015-1040-6.
- [3] Kapageridis I.K., 2015. Variable lag variography using K-means clustering. *Computers & Geosciences*. In Press, Corrected Proof, Available online 13 April 2015.
- [4] Li Z., Kan G., Yao C., Liu Z., Li Q., Yu S., 2014. An improved neural network model and its application in hydrological simulation. *Journal of Hydrologic Engineering*. 19 (10),

04014019-1 – 04014019-17.

- [5] MATLAB, 2012. User's manual. MATLAB documentation.
- [6] Mohammad T. Dastorani, Jamile Salimi Koochi, Hamed Sharifi Darani, Ali Talebi and M. H. Rahimian, 2013. River instantaneous peak flow estimation using daily flow data and machine-learning-based models. *J. Hydroinform.* 245, 1089-1098.
- [7] Piotrowski, A.P., Napiorkowski, J.J., 2013. A comparison of methods to avoid overfitting in neural networks training in the case of catchment runoff modeling. *J. Hydrol.* 476, 97-111.
- [8] Rehab D., Mohammed A., 2015. A novel approach for initializing the spherical K-means clustering algorithm. *Simulation Modelling Practice and Theory.* 54, 49-63.
- [9] Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics.* 20, 53-65.
- [10] Shouke Wei, Depeng Zuo and Jinxi Song, 2012. Improving prediction accuracy of river discharge time series using a Wavelet-NAR artificial neural network. *J. Hydroinform.* 143, 974-991.
- [11] Vahid Nourani and Masoumeh Parhizkar, 2013. Conjunction of SOM-based feature extraction method and hybrid wavelet-ANN approach for rainfall-runoff modeling. *J. Hydroinform.* 141, 829-848.
- [12] Yao, C., Li, Z., Bao, H., Yu, Z., 2009. Application of a developed Grid-Xinjiang model to Chinese watersheds for flood forecasting purpose. *Journal of Hydrologic Engineering*, 14 (9), 923-934.
- [13] Zhao, R., Zhang, Y., Fang, L., Liu, X., Zhang, Q., 1980. The Xinjiang model. In: *Hydrological Forecasting Proceedings Oxford Symposium.* IAHS. pp. 351-356.
- [14] Zhao, R., 1984. *Watershed Hydrological Modeling - Xinjiang Model and Northern Shaanxi Model.* China Water Power Press, Beijing, China, pp. 126-127 (in Chinese).
- [15] Zhao, R., 1992. The Xinjiang model applied in China. *Journal of Hydrology* 135, 371-381.

Acknowledgments

This research was funded by the IWHR Research & Development Support Program (JZ0145B052016) IWHR Scientific Research Projects of Outstanding Young Scientists "Research and application on the fast global optimization method for the Xinjiang model parameters based on the high performance heterogeneous computing" (No. KY1605), Specific Research of China Institute of Water Resources and Hydropower Research (Grant Nos. Fangji 1240), the Third Sub-Project: Flood Forecasting, Controlling and Flood Prevention Aided Software Development - Flood Control Early Warning Communication System and Flood Forecasting, Controlling and Flood Prevention Aided Software Development for Poyang Lake Area of Jiangxi Province (0628-136006104242, JZ0205A432013, SLXMB200902), Study of distributed flood risk forecast model and technology based on multi-source data integration and hydrometeorological coupling system (2013CB036400), and the IWHR application project of multi-source precipitation fusion and soil moisture remote sensing assimilation. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.