Abstract: Machine learning-based data-driven models have achieved great success since their invention. Nowadays, the artificial neural network (ANN)-based machine learning methods have made great progress than ever before, such as the deep learning and reinforcement learning, etc. In this study, we coupled the ANN with the K-nearest neighbor method to propose a novel hybrid machine learning (HML) hydrological model for flood forecast purpose. The advantage of the proposed model over traditional neural network models is that it can predict discharge continuously without accuracy loss owed to its specially designed model structure. In order to overcome the local minimum issue of the traditional neural network training, a genetic algorithm and Levenberg–Marquardt-based multi-objective training method was also proposed. Real-world applications of the HML hydrological model indicated its satisfactory performance and reliable stability, which enlightened the possibility of further applications of the HML hydrological model in flood forecast problems.

Keywords: hybrid machine learning, hydrological model, flood forecast, artificial neural network, K-nearest neighbor method

1 Introduction

Machine learning is a bionics-based problem solution scheme which mimics the functionality of human brain [1–13]. It is usually constituted by a series of mathematic operations and transformations which are implemented by computer programs coupled with modern electronic computers. Among the multitude machine learning methods, the artificial neural networks (ANN) have been paid close attention to by many researchers and engineers for decades. Especially nowadays, with the further improvements of the back-propagation training algorithms and computing horsepower of the modern graphics processing unit (GPU), the ANN technology meets its second spring. Modern ANN technologies, such as the deep learning and the reinforcement learning, etc., have been successfully and widely applied in computer vision, text and speech recognition, self-driving cars, artificial intelligence, and even Go game [14–16].

Many researchers and scientists have tried to apply the ANN technology to solve flood forecast problems and have achieved meaningful results [17–30]. However, due to the high nonlinearity inherent in the flood forecast, the training and testing accuracies usually cannot be both satisfactory which indicate that the forecast capability and stability is not good enough. The poor results are mainly attributed to the following reasons: (a) the network topology and parameters cannot be optimized simultaneously and globally; (b) network with complex topology can satisfactorily simulate in training period, however, the testing results are usually not good. On the contrary, simple network cannot achieve good enough simulations in the training and testing periods. This means that compromise should be made between the network complexity and training accuracy; (c) network designers...
want to further improve the performance of ANN, however, they cannot achieve better performance by using ANN model alone.

In order to solve the above-mentioned three issues, we developed a HML hydrological model and apply it to a real-world hourly rainfall-runoff simulation application. The HML hydrological model is built by coupling ANN runoff estimation with a K-nearest neighbor (KNN) runoff error estimation module to further improve its performance. The advantage of the proposed model over traditional neural network models is that it can predict discharge continuously without accuracy loss owed to its specially designed model structure. In order to overcome the local minimum issue of the traditional neural network training, a novel genetic algorithm (GA) and Levenberg–Marquardt back-propagation (LMBP) algorithm-based multi-objective training method was also proposed to solve the training problems encountered in the ANN applications. Real-world application of the HML hydrological model indicated its satisfactory performance and reliable stability, which enlightened the possibility of the further application of the HML hydrological model in flood forecast problems.

2 Methods

2.1 Study area and data collection

The HML hydrological model was applied in the Tunxi watershed (locates in Anhui province of China) whose catchment area is 2697.76 km². It locates adjacent to the southeast coast of China which is the subtropical monsoon climate zone. The annual mean temperature is 17°C. The winter prevails northwest wind with cold, dry, and sunny weather. The summer is full filled by southeast wind with high temperature, strong sunlight, and moist air. During spring and autumn seasons, the cyclone happens frequently and temperature usually changes between cold and hot. The frontal rain usually happens during spring and summer. The typhoon happens frequently in summer and fall. The monsoon circulation and the main mountain ranges direction obey the orthogonal interaction. The mountains block the northern cold wind and the typhoon. The Tunxi watershed goes downward from west to east. The maximum, minimum, and mean elevations are 1,398, 116, and 380 m, respectively. The relative elevation difference is large. The annual mean rainfall is 1,600 mm and uneven within the year and between different years. 50% of the floods often happen from April to June. 20% of the floods happen from July to September and drought also frequently happens during these months. The runoff changes largely within the year and between different years. The vegetation of the Tunxi watershed is good which mainly includes evergreen coniferous forest, deciduous broad-leaved forest, mixed forest, forest land, grassland, woodland, and pasture crops. The soil type is mainly clay loam.

The rain gauging station, river network, and DEM maps of the Tunxi watershed are demonstrated in Figure 1. The hourly rainfall–runoff simulation of the Tunxi watershed is carried based on 20 flood events for calibration and 9 flood events for validation. The flood events were selected from 1982 to 2002. Eleven rain gauging stations, including Wucheng, Shimen, Zuolong, Dalian, Tunxi, Shangxikou, Rucun, Yixian, Yanqian,
2.2 Identification of the rainfall and antecedent runoff orders

The outlet discharge was simulated or forecasted by using the rainfall and antecedent runoff as model inputs. Rainfall is a key meteorological forcing to generate the runoff. The antecedent runoff functions like a proxy of the watershed soil moisture. With these two factors as model inputs, the discharge time series can be forecasted continuously.

The model inputs are rainfall values at present time step and previous time steps combined with the antecedent runoff values. How many previous time steps’ rainfall and antecedent runoff can be taken as model inputs should be identified and we named the identified number of time steps as the model order. In this research, the orders of rainfall and antecedent runoff are estimated by using the partial mutual information feature selection method [31–33].

2.3 ANN–KNN hybrid runoff and runoff error estimations

The HML hydrological model forecasts basin outlet discharge by using the rainfall and antecedent runoff. The modeling approach is as follows:

\[ \text{EQ}(t) = F_{\text{ANN}}[P(t), \ldots, P(t - n_P), AQ(t - 1), \ldots, AQ(t - n_{AQ})] \]

\[ \text{EE}(t) = F_{\text{KNN}}[\text{EQ}(t), P(t), \ldots, P(t - n_P), AQ(t - 1), \ldots, AQ(t - n_{AQ})] \]

\[ Q(t) = \text{EQ}(t) + \text{EE}(t) \]

where EQ and EE denote estimated discharge and discharge error, respectively; \( F_{\text{ANN}} \) and \( F_{\text{KNN}} \) denote ANN-based discharge estimation and KNN-based discharge error estimation, respectively; \( P \) and \( AQ \) denote rainfall and antecedent runoff, respectively; \( n_P \) and \( n_{AQ} \) denote orders of rainfall and antecedent runoff, respectively; \( Q \) denotes simulated discharge value; \( t \) denotes time step.

The simulation process of the HML hydrological model is as follows: At first, the EQ is estimated by the above equation. And then the EE is calculated. At last, the Q is computed by summing EQ and EE. By using the above-mentioned three equations, the discharge and discharge error are estimated and the final simulated value is obtained by summing these two estimated values together. There is one thing should be noted that the AQ used in the forecast process is not the observed antecedent runoff. We adopted simulated discharges at previous time steps as AQ for flood forecasting. This novel model structure is significantly different from traditional ANN flood forecasting models.

2.4 GA and LMBP-based multi-objective network topology and parameters optimizations

The ANN network topology and parameters are encoded by real numbers as described by Kan et al. [18,19]. These encoded real numbers are treated as decision variables. Therefore, the network topology and parameters can be optimized simultaneously and globally by evolving the decision variables using multi-objective GA (NSGA-II) algorithm. After decoding the evolved decision variables into a set of neural networks, we use the LMBP algorithm to further train the neural networks’ parameters to obtain the final optimal neural networks. The objective functions of the multi-objective optimization are training error and network complexity [18,19]. At last, we choose the network corresponding to the best training error as the final accepted neural network. After the neural network is obtained, the KNN algorithm is calibrated by using the leave-one-out cross-validation method [18].

3 Results and discussion

3.1 Runoff scatter plot analysis

The scatter plots of the observed and simulated discharges of the HML hydrological model are demonstrated in Figure 2. As we can see from Figure 2(a) and (b), the calibration and validation results are satisfactorily good. The regression \( R^2 \) values of the calibration and validation periods are 0.9635 and 0.9515, respectively. The scatter plot of the calibration period demonstrates a good and even distribution. The scatter plot of the validation period is a bit worse than the calibration period.
The scatters are dispersed and not concentrated along with the 45° line. The regression $R^2$ value of the validation period (0.9515) is also worse than the calibration period (0.9635). Although the simulation result of the validation period is not as good as the calibration period, the HML hydrological model's performance in validation period is also good enough. The accuracy decrement ratio from the calibration period to the validation period is $0.9515/0.9635 \approx 0.9875$ which is acceptable for real-world applications. It also can be found in Figure 2(b) that the simulation results of validation period are biased (underestimated). This phenomenon will be further discussed in the following error statistics sections.

3.2 Hydrograph analysis

The hydrographs simulated by the HML hydrological model are accurate and smooth. The satisfactory simulation property of the HML hydrological model is owed to the modeling approach and the proposed model calibration method. The good accuracy of the simulation results is owed to the multistep rainfall input variables. With continuous rainfall inputs, the water balance can be ensured. The ANN–KNN hybrid data-driven method can make full use of rainfall input information and simulates the flood accurately. Therefore, the simulation is accurate and stable. The good smoothness of the simulated hydrographs is related to the antecedent runoff values. The continuously antecedent runoff inputs ensure the continuity of the discharge time series and make the simulated hydrographs very smooth. For the reason of space limitation, we only demonstrate four typical hydrographs selected from the calibration and validation periods in Figure 3.

3.3 Error statistics of flood events

The error statistics of flood events simulations are listed in Table 1. As we can see in the table, the calibration and validation accuracies seem very satisfactory. The observed peak flows range from 633 to 6,490 m³/s. The observed total rainfall ranges from 74.2 to 1199.9 mm. The maximum and minimum peak flows of the calibration period are 4,700 and 633 m³/s, respectively. The maximum and minimum peak flows of the validation period are 6,490 and 1,410 m³/s, respectively. The maximum peak flow of the validation period exceeds the peak flow range of the calibration period. This may lead to a worse validation performance compared with the calibration period. Detailed analysis of other error statistic criteria can be found in the following paragraphs.

3.3.1 Total volume relative error

As for the calibration period, there are 14 negative relative errors and 6 positive relative errors. This result indicates that the model calibrations are preferred to underestimate the total volume of the runoff. As for the validation period, there are four negative relative errors and five positive relative errors. This result shows that the validation period do not overestimate or underestimate the total volume significantly. In the calibration period, only flood event 1987061908's absolute total volume relative error exceeds
20%. These results prove that the HML hydrological model performs well in total volume estimations, especially in the validation period. The HML hydrological model has good generalization capability and obtains good results in the validation period.

### 3.3.2 Peak flow relative error

The peak flow relative errors of calibration period range from −30.3% to 59.6%. There are three flood events’ absolute peak flow relative errors exceed 20%. They are 1987050108, 1987061908, and 1991063008. As for the validation period, no flood event’s absolute peak flow relative error exceeds 20%. As for the calibration period, there are 11 negative relative errors and 9 positive relative errors. As for the validation period, there are six negative relative errors and three positive relative errors. These results indicate that the peak flow simulations are not significantly underestimated or overestimated in the calibration period, however, they are underestimated in the validation period for six flood events. As for the flood event 1996060100, which has the maximum peak flow value throughout the calibration and validation period, the HML hydrological model successfully forecasts the peak value within the 20% error threshold even it exceeds the maximum peak flow value of the calibration period. This result proves the excellent generalization capability of the HML hydrological model.

### 3.3.3 Peak flow present time error

There are two flood events exceeding the peak flow present time error threshold (3 h) in the calibration period. There is only one flood event exceeding the peak flow present time error threshold (3 h) in the validation period. Generally speaking, the HML hydrological model performs well in the peak flow present time estimation. We also notice that, as for the calibration period, there are five negative errors and eight positive errors and as for the validation period, there are two negative errors and four positive errors. The HML hydrological model seems inclined to forecast peak flow a bit ahead of time.

### 3.3.4 Nash–Sutcliffe coefficient of efficiency

The HML hydrological model performs very well considering the criterion of Nash–Sutcliffe coefficient of efficiency (NSCE). Most flood events achieve NSCE higher
or equal to 0.9. Only flood event 1987050108 and 1987061908 obtain NSCE of 0.52 and 0.79, respectively. We notice that these two events’ peak flows are relatively small (633 and 945 m³/s) compared with other flood events. This may be the reason why the HML hydrological model performs not well in these two events’ simulations.

<table>
<thead>
<tr>
<th>Flood ID</th>
<th>Calibration or validation periods</th>
<th>Total rainfall (mm)</th>
<th>Observed peak flow (m³/s)</th>
<th>Total volume relative error (%)</th>
<th>Peak flow relative error (%)</th>
<th>Peak flow present time error (h)</th>
<th>Nash–Sutcliffe coefficient of efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1982050108</td>
<td>Calibration</td>
<td>512.0</td>
<td>4,280</td>
<td>-0.08</td>
<td>-0.70</td>
<td>-1</td>
<td>0.98</td>
</tr>
<tr>
<td>1983051108</td>
<td>75.5</td>
<td>1,300</td>
<td>1.81</td>
<td>3.20</td>
<td>0</td>
<td>0.96</td>
<td></td>
</tr>
<tr>
<td>1983051422</td>
<td>79.9</td>
<td>1,510</td>
<td>-6.94</td>
<td>8.30</td>
<td>0</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>1983060906</td>
<td>119.4</td>
<td>2,170</td>
<td>-0.38</td>
<td>-8.70</td>
<td>3</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>1984082620</td>
<td>172.7</td>
<td>2,513</td>
<td>1.99</td>
<td>-4.50</td>
<td>-4</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>1986061108</td>
<td>511.6</td>
<td>2,260</td>
<td>-3.10</td>
<td>-8.20</td>
<td>0</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>1987050108</td>
<td>74.2</td>
<td>633</td>
<td>7.41</td>
<td>59.60</td>
<td>-1</td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td>1987061908</td>
<td>110.4</td>
<td>945</td>
<td>-22.36</td>
<td>-21.10</td>
<td>-6</td>
<td>0.79</td>
<td></td>
</tr>
<tr>
<td>1988050704</td>
<td>154.6</td>
<td>1,390</td>
<td>3.68</td>
<td>9.50</td>
<td>1</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>1988061101</td>
<td>191.9</td>
<td>1,000</td>
<td>-10.59</td>
<td>-8.20</td>
<td>1</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td>1989050108</td>
<td>312.9</td>
<td>1,740</td>
<td>-12.42</td>
<td>-7.00</td>
<td>0</td>
<td>0.97</td>
<td></td>
</tr>
<tr>
<td>1989061206</td>
<td>197.9</td>
<td>2,274</td>
<td>-3.88</td>
<td>7.40</td>
<td>0</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>1989063023</td>
<td>161.6</td>
<td>1,740</td>
<td>-4.65</td>
<td>-6.90</td>
<td>0</td>
<td>0.97</td>
<td></td>
</tr>
<tr>
<td>1989072208</td>
<td>132.0</td>
<td>1,470</td>
<td>3.88</td>
<td>-5.60</td>
<td>2</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>1990050108</td>
<td>89.5</td>
<td>1,700</td>
<td>7.65</td>
<td>-18.40</td>
<td>1</td>
<td>0.96</td>
<td></td>
</tr>
<tr>
<td>1990061108</td>
<td>355.5</td>
<td>2,500</td>
<td>-7.37</td>
<td>9.70</td>
<td>-3</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>1991051800</td>
<td>504.2</td>
<td>2,220</td>
<td>-10.70</td>
<td>19.70</td>
<td>3</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>1991063008</td>
<td>355.1</td>
<td>2,060</td>
<td>-4.58</td>
<td>-30.30</td>
<td>1</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td>1992062000</td>
<td>357.5</td>
<td>3,150</td>
<td>-10.78</td>
<td>0.70</td>
<td>1</td>
<td>0.96</td>
<td></td>
</tr>
<tr>
<td>1993052700</td>
<td>1074.4</td>
<td>4,700</td>
<td>-2.79</td>
<td>18.30</td>
<td>0</td>
<td>0.96</td>
<td></td>
</tr>
<tr>
<td>1994050100</td>
<td>Validation</td>
<td>779.0</td>
<td>4,160</td>
<td>-9.16</td>
<td>-0.30</td>
<td>1</td>
<td>0.97</td>
</tr>
<tr>
<td>1995051500</td>
<td>1010.9</td>
<td>4,070</td>
<td>-7.97</td>
<td>-13.70</td>
<td>3</td>
<td>0.89</td>
<td></td>
</tr>
<tr>
<td>1996060100</td>
<td>1199.9</td>
<td>6,490</td>
<td>-5.35</td>
<td>-14.10</td>
<td>-1</td>
<td>0.96</td>
<td></td>
</tr>
<tr>
<td>1997060600</td>
<td>552.5</td>
<td>2,730</td>
<td>3.81</td>
<td>-3.50</td>
<td>3</td>
<td>0.96</td>
<td></td>
</tr>
<tr>
<td>1999052108</td>
<td>208.4</td>
<td>2,960</td>
<td>0.10</td>
<td>-6.10</td>
<td>-1</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>1999062215</td>
<td>617.0</td>
<td>3,780</td>
<td>-14.47</td>
<td>-10.20</td>
<td>0</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>2001050108</td>
<td>98.1</td>
<td>1,410</td>
<td>24.24</td>
<td>18.10</td>
<td>0</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>2001062008</td>
<td>297.2</td>
<td>3,640</td>
<td>7.09</td>
<td>0.50</td>
<td>6</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>2002051308</td>
<td>536.2</td>
<td>2,120</td>
<td>5.89</td>
<td>0.40</td>
<td>0</td>
<td>0.89</td>
<td></td>
</tr>
</tbody>
</table>

4 Conclusions

In this study, we coupled the ANN with KNN method to build a novel HML hydrological model for flood forecast purpose. The advantage of the proposed model over traditional neural network models is that it can predict discharge continuously without accuracy loss owed to its specially designed model structure. In order to overcome the local minimum issue of the network training, a GA and Levenberg–Marquardt-based multi-objective training method was also proposed. Real-world application of the HML hydrological model indicated its satisfactory performance and reliable stability, which enlightened the possibility of the further application of the HML hydrological model in flood forecast problems.

**Acknowledgments:** This research was funded by Beijing Natural Science Foundation (8181001), SKL of HESS-1914, National Natural Science Foundation of China (51909272), National Key R&D Program of China (2017YFC1502706, 2019YFC1510605), and IWHR Research & Development Support Program (JZ0145B022017). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 and TITAN V GPU used for this research. Guangyuan Kan and Ke Liang are corresponding authors. The authors declare that there is no conflict of interest regarding the publication of this paper.
**Author contribution:** Guangyuan Kan designed the experiments and Ke Liang carried them out. Guangyuan Kan developed the model code and Ke Liang performed the simulations. Guangyuan Kan and Ke Liang prepared the manuscript with contributions from all co-authors. The authors applied the SDC approach for the sequence of authors.

**References**


