Wavelet-genetic programming conjunction model for flood forecasting in rivers

Mani Kumar and Rajeev Ranjan Sahay

ABSTRACT

In this study we have developed a conjunction model, WGP, of discrete wavelet transform (DWT) and genetic programming (GP) for forecasting river floods when the only data available are the historical daily flows. DWT is used for denoising and smoothening the observed flow time series on which GP is implemented to get the next-day flood. The new model is compared with autoregressive (AR) and stand-alone GP models. All models are calibrated and tested on the Kosi River which is one of the most devastating rivers of the world with high and spiky monsoon flows, modeling of which poses a great challenge. With different inputs, 12 models, four in each class of WGP, GP and AR, are devised. The best performing WGP model, WGP4, with four previous daily flow rates as input, forecasts the Kosi floods with an accuracy of 87.9%, root mean square error of 123.9 m^3/s and Nash–Sutcliffe coefficient of 0.993, the best performance indices among all the developed models. The extreme floods are also better simulated by the WGP models than by AR and GP models.

Key words | flood modeling, genetic programming, India, Kosi, wavelet transform

Mani Kumar
Rajeev Ranjan Sahay (corresponding author)
Civil Engineering Department,
Birla Institute of Technology (Mesra),
Patna 800014,
India
E-mail: *rajeev_sahay@yahoo.com*

INTRODUCTION

Floods kill more people and destroy more property every year than any other type of natural disaster. According to the estimate by the World Resources Institute, India has a larger percentage of population exposed to flood damage than any other country of the world. Bangladesh, China, Vietnam, Pakistan and Indonesia are other countries with significant percentages of their population at flood risk. Additionally, it is worrying that climatic change is increasing the frequency and severity of extreme events (Li *et al.* 2016). However, flood damage can be considerably reduced if their occurrence is predicted reliably in advance. Authorities require a timely and reliable estimate of any impending deluge to devise appropriate action plans. Numerous models developed for flood forecasting in the last decades broadly employ three approaches: conceptual, statistical, or artificial intelligence. Although the conceptual approach has an apparent physical concept and consistent accuracy, its heavy hydrometeorological data requirement is a deterrent to its use. The statistical approach, on the other hand, is simple to formulate and execute, but does not give satisfactory results as it is static and deficient in revealing the nonlinear relationship that may exist between the affecting variables. In comparison, models based on artificial intelligence, such as artificial neural networks (ANN), have the ability to learn the fluctuating relationship between the input and output without knowledge of the physical processes occurring within the system; however, their inability to produce an explicit model for use of other investigators is a clear disadvantage. In addition, their optimal structures are hard to determine and the network solutions are easily trapped in local optima.

Genetic programming (GP), however, is a recently developed evolutionary computing technique which generates a transparent and structured representation of the

doi: 10.2166/nh.2018.183

studied system (Koza 1992). It is free from the deficiencies of ANN and provides an explicit model. These advantages are successfully applied in hydrological investigations (Savic *et al.* 1999; Babovic & Keijzer 2002; Dorado *et al.* 2003; Aytek & Kisi 2008; Nourani *et al.* 2008; Azamathulla *et al.* 2010; Shiri & Kisi 2010; Kisi & Shiri 2012; Shiri *et al.* 2012a, 2012b; Fallah-Mehdipour *et al.* 2012; Rodríguez-Vázquez *et al.* 2012; Fallah-Mehdipour *et al.* 2013; Kisi *et al.* 2013; Ashofteh *et al.* 2014; Aytek *et al.* 2014; Garg 2014; Hakimzadeh *et al.* 2014; Uyumaz *et al.* 2014; Yaseen *et al.* 2015; Adhikary *et al.* 2016; Barge & Sharif 2016; Karimi *et al.* 2017). However, when the time series is highly fluctuating and nonstationary, preprocessing of the series through a suitable technique is desirable for better predictability (Cannas *et al.* 2006). The discrete wavelet transform (DWT) is an important signal decomposition technique that extracts valuable characteristics of signals at different resolution levels and which, when combined with an evolutionary algorithm such as GP, provides an efficient nonlinear approximation model. Very little literature is available on the combined use of DWT and GP in hydrology. Kisi & Shiri (2011) formulated WGP for daily precipitation forecasting. Shiri & Kisi (2012) modeled daily sediment load and showed that it performed better than wavelet-neuro-fuzzy and wavelet-ANN models. Nourani *et al.* (2012) developed a hybrid model of wavelet transform and GP to optimize the ANN structure for rainfall-runoff forecasting and the results thus obtained were found to be better than ANN and GP models. Gorgij *et al.* (2017) developed a WGP model for groundwater budget forecasting in East Azerbaijan, Iran. Karimi *et al.* (2016) implemented WGP for flow forecasting in the small River Filyos in Turkey and showed it to be more efficient than autoregressive moving average and ANN models.

In the present work, a novel model combining DWT and GP is developed for predicting the next-day Kosi flows for the monsoon period at Baltara gauge site (India) where the only records available are the previous day's flow rates. The motivation for using the WGP conjunction model arises from the fact that the wavelet analysis captures the nonstationary and seasonal effects of the flow time series and the input parameters are minimized based on sensitivity analysis through the GP framework (Nourani *et al.* 2012). As mentioned above, only limited literature is available in which wavelet transform has been combined with GP for stream flow forecasting in small rivers. However, this should be perhaps the first time that WGP has been evaluated for flood forecasting in a large river like the Kosi, which has a highly undulating monsoon flow that is marked by drifts, trends and abrupt changes.

MODELS

Autoregression

The stream flow on any day is correlated with the flows on preceding days. The order of an autoregressive model is the number of immediately preceding values in the time series that are used to predict the current value. An n^{th} order autoregressive model, AR*n*, is represented by:

$$y_i = \sum_{i=1}^{n} B_i y_{i-1} + \varepsilon_t \tag{1}$$

where B_i are the autoregression coefficients evaluated using the least squares method, y_i is the time series under investigation and ε_t is the residue term which is assumed to be the Gaussian white noise.

Discrete wavelet transform

The Fourier transform (FT) and wavelet transform (WT) are widely used techniques for analyzing time series. But as FT loses time information while dissociating the series into frequency domain, it is found to be unsuitable for analyzing signals with transitory characteristics (Partal & Kucuk 2006; Partal & Kisi 2007). WT, in comparison, resolves both time and scale events better. WT dissociates a time series into the translated and dilated/compressed version of a suitable mother wavelet and generates wavelet coefficients $C_{a,b}$ at different steps which measure the correlation of the wavelet to the time series. This process, called the continuous wavelet transform of the signal *f(t)* w.r.t. the mother wavelet/basis function $\varphi(t)$, is given by

$$C_{a,b} = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} f(t).\ \varphi^*\left(\frac{t-b}{a}\right) dt \tag{2}$$

The conjugate functions $\varphi*(t-b/a)$ are derived by scaling the mother wavelet by '*a*' and translating it by '*b*'.

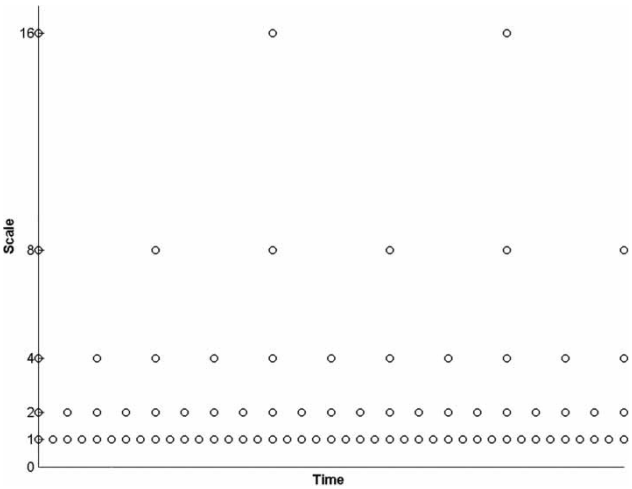

Figure 1 | The dyadic scheme of discrete time series decomposition.

However, as flow in a river is measured at intervals and hence discontinuously, a DWT instead of continuous wavelet transform is more suitable. In addition, DWT alleviates the computational burden by adopting the dyadic scheme in which at every step, the scale and step spacing between wavelets is increased by a factor of two, i.e. in the first pass wavelets of unit scale with unit space between them are chosen, while in the second pass wavelets of double scale and double space between them are chosen, and so on. The dyadic scheme of discrete time series decomposition is illustrated in Figure 1. Thus, for a discrete time series *x(t)*, the dyadic DWT is given by

$$T_{a,b} = 2^{-a/2} \sum_{t=0}^{N-1} x(t)\, \varphi^{(2^{-a}t-b)} \tag{3}$$

where $T_{a,b}$ is the coefficient for the DWT and N is the signal length.

In this work, the time series is decomposed using the Mallat (1989) filtering process in which the time series is passed through the low-pass and high-pass filters simultaneously. The outcomes through the low-pass filters are the high-scale and low-frequency wavelets. They are called the approximation wavelets/subtime series, A_i, at resolution level i. They represent gross and slow-changing features of the signal. On the other hand, the outcomes through the high-pass filters are low-scale and high-frequency wavelets, called the detail wavelets/subtime series, D_i. They measure the rapidly changing features of the signal. The approximation subtime series can be further decomposed into approximation and detail wavelets for the next decomposition/resolution level. A three-level wavelet decomposition is illustrated in Figure 2.

The original signal can again be regained by adding all the detail subtime series and the approximation subtime series of the last resolution level.

Genetic programming

GP is a comparatively new machine learning method that randomly generates a population of computer programs. These programs are potential candidates for solution. The fitness, i.e., suitability of each generated candidate for solution, is evaluated. The 'fit' candidates are kept and 'misfit' candidates are removed from the next generation. In addition, some new candidates are formed randomly by changing (mutation) or swapping parts of the other

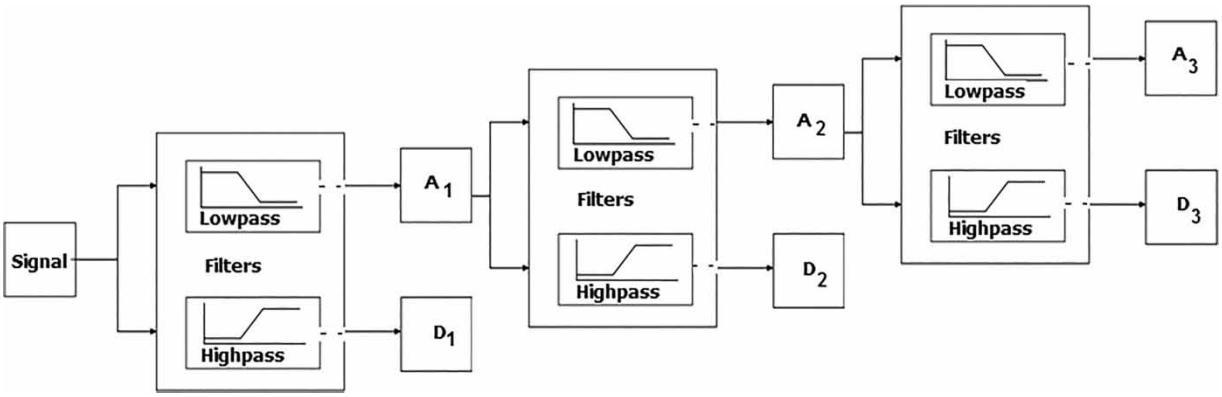

Figure 2 | Mallat (1989) method of time series decomposition into approximation and details subtime series.

candidates (crossover). The process is repeated until the generated programs of the newer generation meet the objective of minimizing the root mean square error between the observed and the predicted flows for the derivation dataset. Although GP and genetic algorithms (GA) have similar working structure, GP has some distinct advantages over GA. While GA has a fixed length, meaning the resulting function has bounded complexity, GP inherently has a variable length making it more flexible with no bounded complexity. Furthermore, GA relies on operator precedence which could be seen as a limitation, whereas GP uses an explicit structure to avoid operator precedence (Kisi & Shiri 2011). The books by Koza (1992) and Banzhaf *et al.* (1998) have given a good description of GP.

Discrete wavelet transform – genetic programming (WGP)

The WGP is an integration of DWT and GP. DWT is used for denoising the flow time series. This is done by decomposing the series by DWT into approximation and detail subtime series. Generally, the detail subtime series on the first resolution level, i.e., D_1, is found to be the noisiest and least correlated subtime series with the original series (Kisi & Shiri 2012). Therefore, by deleting D_1 and recombining other wavelets, a wavelet-smoothened series is formed. This transformed series is the input for the GP implementation. The working arrangement for WGP is shown in Figure 3.

DATA

Kosi river

The monsoon floods are a recurring hazard in Eastern India. The Kosi river, in particular, causes widespread human suffering every year through flooding. It is a transboundary Himalayan river with the major portion of its catchment lying in China and Nepal. Although only about 11,410 km^2 out of its total catchment area of 74,030 km^2 lies in India, it bears the brunt of the Kosi's fury. Any heavy rainfall in its upper catchment results in high floods which rush towards the Indian plain and take a heavy toll on life and property. Additionally, its steep slope in the gorge zone causes high flow velocity which erodes the fragile Himalayan topsoil resulting in a silt yield of about 19 m^3/ha/year, one of the highest in the world. The silt is released when the flow velocity diminishes in the plain, causing frequent shifts in the river's course. It has moved laterally westward by about 114 km in the last 250 years. To control the flood damage and confine its course, the government of the state of Bihar has constructed earthen embankments on both sides of the river, without much satisfactory result. The embankments are often breached or overtopped, submerging a large area. The disaster happens almost every year. In the year 2008, the embankment breach was so catastrophic that the government had to declare it a national calamity and a large contingent of the army had to be called in to evacuate the affected people. The Kosi flow increases manyfold during monsoon as many big rivers such as the Kamala, Bagmati and Bhutahi Balan flow into it. The Kosi catchment along with its adjoining rivers is shown in Figure 4.

In this research, a MATLAB-based open-source genetic-programming tool, GPTIPS (Searson *et al.* 2010) is used on the wavelet-smoothened time series of the Kosi river at Baltara to predict the next-day flows. Different GP parameters, i.e., population size, number of generations, tournament size, elitism, depth of tree, number of genes in a tree and the allowed mathematical functions are tried in every run. The objective for each GP-run is to minimize the root mean square error between the observed and the predicted flows on the derivation dataset. The derivation dataset consisted of daily discharges of the monsoon period (June–October) for the years 2001–4, 2006–7 and 2009–12. The discharges for the years 2005 and 2008 could not be considered, as the river breached its banks in these years and formed two streams. After deriving the proposed models satisfactorily using the derivation dataset, they are tested on the verification dataset which consisted of monsoon discharges for the years 2013–16. The statistical information on the Kosi flow at Baltara is shown in Table 1.

As the first step, the observed flow time series (OFTS) of the Kosi is dissociated into wavelets /subtime series at three decomposition levels (Figure 5). Wang & Ding (2003) suggested int[log(n)] decomposition levels sufficient for bringing out the attributes of the signal, where n is the data length. As only 1,137 daily flow data are used for

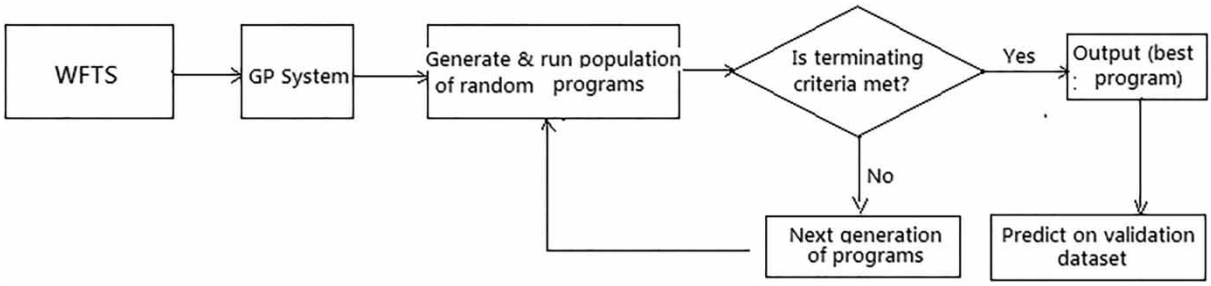

Figure 3 | The working structure of the WGP model.

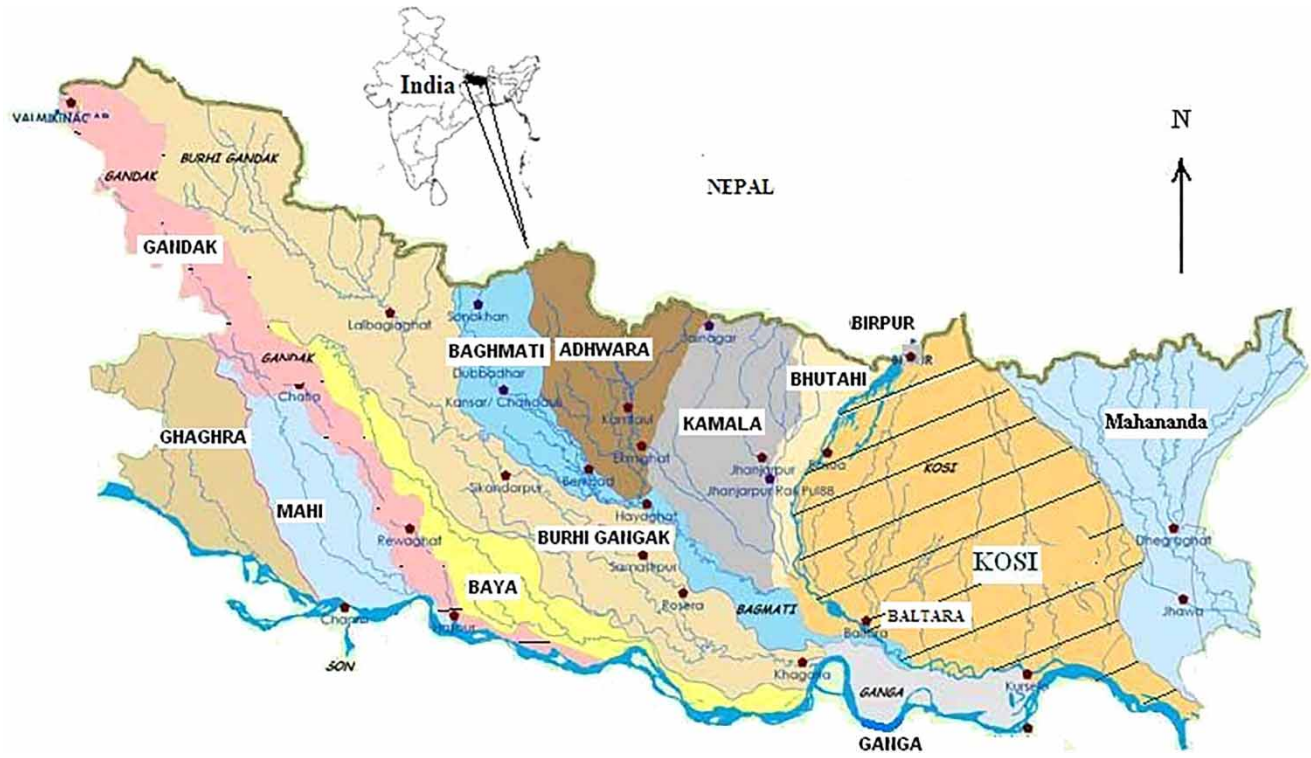

Figure 4 | The catchment of the Kosi and the adjacent rivers (FMIS 2016).

deriving the models, three decomposition levels are considered adequate. Furthermore, a suitable mother wavelet is critical to the efficiency of a wavelet-based model. Some of the popular wavelets used in hydrological studies are haar, coif5, dmey, bior6.5, rbio6.8, sym9 and db5. In the present study, rbio6.8 is found to be the most appropriate wavelet for preprocessing of the Kosi flows at Baltara, as the prediction based on it has the maximum correlation coefficient and the minimum root mean square error, both for the derivation and the verification datasets (Table 2).

Table 1 | The flow characteristics for the Kosi at Baltara (2002–16)

Parameter (m^3/s)	Derivation dataset	Verification dataset
Max. daily discharge	7,265	6,268
Min. daily discharge	866	653
Mean daily discharge	4,255	3,905
Standard deviation	1,367	1,293
Range	6,399	5,615

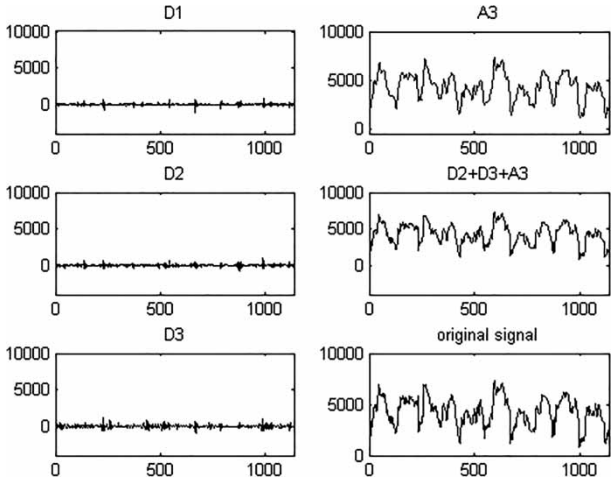

Figure 5 | The Kosi flow decomposition into approximation and detail wavelets at three resolution levels.

Again, a correlation study between the subtime series and the original flow time series showed D_1, the first detail subtime series is found to be the noisiest and the least correlated wavelet to the original flow time series. Therefore, ignoring D_1 and re-adding wavelets D_2, D_3 and A_3, a wavelet-smoothened flow time series (WFTS) is developed which formed inputs for WGP models. The AR and GP models use OFTS for inputs. Based on different inputs, 12 models (four in each class of WGP, GP and AR) are developed, i.e., models WGP1, WGP2, WGP3 and WGP4 in the WGP class, models GP1, GP2, GP3 and GP4 in the GP class, and models AR1, AR2, AR3 and AR4 in the AR class. The model WGP3, for instance, has three inputs, i.e., the current-day, one-day-before and two-days-before flow rates from WFTS, while models GP3 and AR3 have inputs from OFTS (Table 3). The desired output for all the proposed models is the next-day flow.

Table 2 | Performance of the WGP4 model with different mother wavelets. NSC, Nash–Sutcliffe coefficient; RSME, root mean square error

	Derivation period		Verification period	
Mother wavelet	**NSC**	**RMSE (m^3/s)**	**NSC**	**RMSE (m^3/s)**
coif5	0.983	161	0.973	159
dmey	0.956	172	0.963	169
bior6.8	0.949	181	0.950	178
rbio6.8	0.989	155	0.992	124
sym9	0.957	170	0.953	174
db5	0.986	159	0.980	160
haar	0.883	196	0.869	191

Table 3 | Input variables of the proposed models

Model	Inputs	Target
AR1/GP1	$Q_{o,t}$	$Q_{o,t+1}$
AR2/GP2	$Q_{o,t}$, $Q_{o,t-1}$	$Q_{o,t+1}$
AR3/GP3	$Q_{o,t}$, $Q_{o,t-1}$, $Q_{o,t-2}$	$Q_{o,t+1}$
AR4/GP4	$Q_{o,t}$, $Q_{o,t-1}$, $Q_{o,t-2}$, $Q_{o,t-3}$	$Q_{o,t+1}$
WGP1	q_t	$Q_{o,t+1}$
WGP2	q_t, q_{t-1}	$Q_{o,t+1}$
WGP3	q_t, q_{t-1}, q_{t-2}	$Q_{o,t+1}$
WGP4	q_t, q_{t-1}, q_{t-2}, q_{t-3}	$Q_{o,t+1}$

Note: $Q_{o,t-i}$ and $q_{t-i,\ (i=0\ \text{to}\ 3)}$ are i-day-before flows from OFTS and WFTS, respectively, while $Q_{o,t+1}$ is the next-day flow from OFTS.

After models are satisfactorily derived, they are evaluated using the following performance indices:

(i) Nash–Sutcliffe coefficient,

$$\text{NSC} = 1 - \frac{\sum_{t=1}^{N} (Q_{p,t} - Q_{o,t})^2}{\sum_{t=1}^{N} (Q_{o,t} - \overline{Q}_{o,t})^2} \tag{4}$$

(ii) Root mean square error,

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{N} (Q_{p,t} - Q_{o,t})^2}{N}} \tag{5}$$

$$\text{(iii) \% Accuracy} = \frac{100\ N'}{N} \tag{6}$$

where $Q_{p,t}$ and $Q_{o,t}$ are the predicted and observed flow rates on any day 't', $\overline{Q}_{o,t}$ is the mean of the observed flows, N is the data length and N' is the number of forecast flows with maximum deviation up to 5% from the respective observed flows.

RESULTS

The proposed AR, GP and WGP models are derived and verified for forecasting of the next-day Kosi floods for the

Table 4 | The performance of the proposed models for the Kosi River at Baltara (India)

		Derivation dataset			Verification dataset		
Set	**Model**	**NSC**	**RMSE (m^3/s)**	**Accuracy (%)**	**NSC**	**RMSE (m^3/s)**	**Accuracy (%)**
Set 1	AR1	0.96	258	78.0	0.97	214	70.5
	GP1	0.96	257	77.8	0.97	215	69.9
	WGP1	0.97	227	79.4	0.98	185	74.6
Set 2	AR2	0.97	249	81.1	0.97	207	72.7
	GP2	0.97	249	81.0	0.97	207	72.9
	WGP2	0.98	198	83.2	0.99	159	79.4
Set 3	AR3	0.97	249	81.1	0.97	207	72.7
	GP3	0.96	246	81.2	0.97	206	73.0
	WGP3	0.98	175	86.9	0.99	135	77.4
Set 4	AR4	0.97	248	81.0	0.97	207	72.7
	GP4	0.97	245	81.4	0.97	206	73.1
	WGP4	0.99	155	88.4	0.99	124	87.9

monsoon period (June–October) at the Baltara gauge site. The performance of the devised models as presented in Table 4 clearly suggests that WGP is superior to GP and AR models. This is because the WGP model captures long, intermediate and short patterns in the flow time series, whereas GP and AR capture short patterns only (Nourani *et al.* 2012; Papacharalampous *et al.* 2017).

In order to illustrate the comparison better, the models are grouped into four sets. Set 1 consists of three models: AR1, GP1 and WGP1. They are simple models with just one input, the current-day flow, either from OFTS (for GP1 and AR1) or from WFTS (for WGP1). As can be observed from Table 4, the WGP1 (Equation (7)) is found to be a reasonably good model with 74.6% prediction accuracy for the verification dataset, while the corresponding accuracies for AR1 (Equation (8)) and GP1 (Equation (9)) are only 70.5% and 69.9% respectively. The model WGP1 also fares better on the other performance indices, i.e., NSC and RMSE. This is true for the derivation dataset as well. It must be mentioned here again that a prediction is considered accurate if it does not deviate from the observed value by more than 5%.

$$Q_{p,\,t+1} = 124.5 + 0.948q_t + 0.48 \times 10^{-5} q_t^2 \tag{7}$$

$$Q_{p,\,t+1} = 80.2 + 0.98Q_{o,t} \tag{8}$$

$$Q_{p,\,t+1} = 186.9 + 0.923\mathrm{Q_{o,t}} + 0.71 \times 10^{-5} Q_{o,t}^2 \tag{9}$$

where $Q_{o,t}$ and $Q_{p,\,t+1}$ are the observed and the predicted flows on days 't' and '$t+1$' respectively, and q_t is the WFTS flow on day 't'.

The models in Set 2, i.e., AR2, GP2 and WGP2, have an additional input, the previous-day flow-rate, in addition to the present-day flowrate. Evidently, performance of all models improves. The model WGP2, with its NSC increasing from 0.98 to 0.99, RMSE decreasing from 185 m^3/s to 159 m^3/s and accuracy increasing to 79.4% for the verification dataset, is again found to be better than the AR2 and GP2 models. For the derivation dataset also, WGP2 outperforms AR2 and GP2. Models WGP3, GP3 and AR3 in Set 3 have three previous days' flow-rates as inputs. However, the additional input hardly makes any improvement in the performance of the models. The concentration time for the Kosi at Baltara as estimated by the Kirpich formula is 4.31 days, suggesting that the flood water takes as many days from the farthest place in the basin to reach the gauge site during the monsoon period. Therefore, models WGP4, GP4 and AR4 are developed with four preceding daily flow-rates as the input. After trying several GP runs, the optimal WGP4 model (Equation (10)) is obtained with the following GP parameters: population size = 200, number of generation = 200, tournament size = 4, elitism = 0.01% of population, maximum depth of tree = 3, number of genes allowed in a tree = 4 and function node set = {plus, minus, times, power}. The model seems to be well derived as its forecast for the derivation dataset has an NSC value of 0.99,

RMSE value of 155 m^3/s and an accuracy value of 88.4%, which is superior to all the developed models. When compared, the models AR4 (Equation (11)) and GP4 (Equation (12)) lag behind in their performance.

$$Q_{p,\,t+1} = 38 + \frac{8290}{q_t} - 3.631\,q_{t-1} + 3.102\,q_{t-2} - 0.351\,q_{t-3} + \left(3.425 - \frac{1.203 q_{t-2}}{q_{t-1}}\right) q_t - 0.3508\,q_{t-2}^2/q_{t-1} \quad (10)$$

$$Q_{p,\,t+1} = 107.9 + 1.23Q_{o,t} - 0.275Q_{o,\,t-1} + 0.103Q_{o,\,t-2} - 0.086Q_{o,\,t-3} \quad (11)$$

$$Q_{p,\,t+1} = 108 - 0.275Q_{o,t} + 0.103Q_{o,\,t-1} + 1.232Q_{o,\,t-2} - 0.09Q_{o,\,t-3} \quad (12)$$

where q_{t-i} and $Q_{o,\,t-i}$, and $i = 0$ to 3, are $(t\text{-}i)^{\text{th}}$ day flow-rates on WFTS and OFTS respectively.

Figure 6 shows the comparison of the high floods (>6,000 m^3/s) predicted for the derivation dataset by the best models of their class, i.e., models WGP4, GP4 and AR4. The WGP4 is again seen to be the most reliable for forecasting the high flows, as the predicted series almost resembles the observed high flows of the Kosi, whereas models AR4 and GP4 show greater deviation. Once the models in Set 4 are satisfactorily derived, they are tested for the verification dataset. It can be seen from Table 4 that in this dataset also the WGP4 is clearly the best forecasting model. It predicts the floods with the highest NSC value of 0.99, the least RMSE value of 124 m^3/s and the maximum accuracy of 87.9%.

Figure 7 compares the predicted and the observed high floods (flows >5,500 m^3/s) for the verification dataset. It can be observed from Table 4 that WGP4 predicts the high flows better than the GP4 and AR4 models. Its predictions for the highest three floods (of 6,268 m^3/s, 6,203 m^3/s and 6,181 m^3/s) as 6,228 m^3/s, 6,214 m^3/s and 6,179 m^3/s respectively, are closest to the observed values, while models AR4 and GP4 either under- or over-predict these floods as 6,160 m^3/s, 6,233 m^3/s and 6,029 m^3/s respectively, and 6,150 m^3/s, 6,222 m^3/s and 6,019 m^3/s respectively. Understandably, there is prevalence of more noise in high discharges which are not normally measured directly but extrapolated from the stage-discharge rating curves (Tiwari & Chatterjee 2011). All models perform similarily in case of low flows which are influenced more by the moisture and vegetative condition of the soil.

The summarized results show that the WGP model is suitable for flood forecasting even in large and unpredictable rivers like the Kosi, and outperforms the AR and GP models. The AR and GP models, in comparison, have modeling abilities which are in agreement with the findings of other

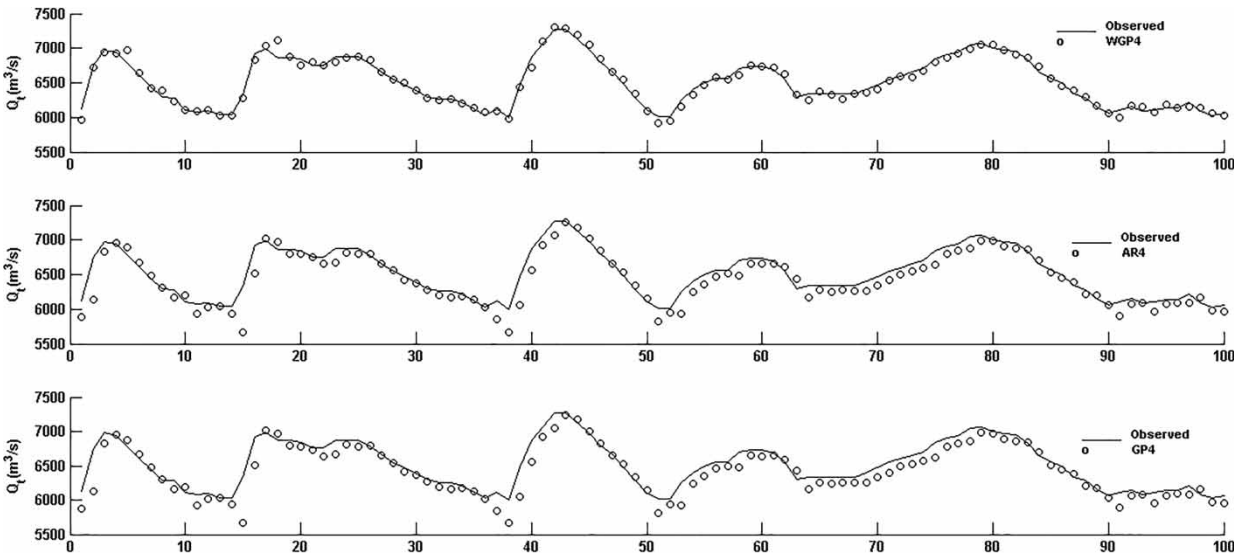

Figure 6 | The comparison of the predicted and observed high flows (>6,000 m^3/s) of the Kosi at Baltara (derivation dataset).

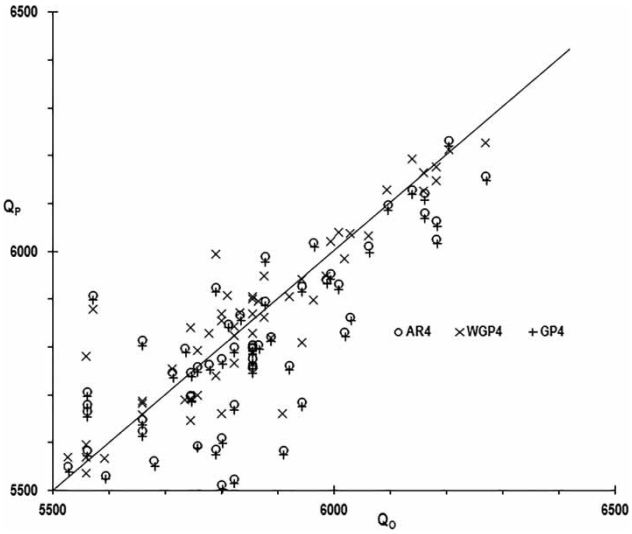

Figure 7 | The comparison of the predicted and observed high flows (>5,500 m^3/s) of the Kosi at Baltara (verification dataset).

investigators (Savic *et al.* 1999; Nourani *et al.* 2012). By combining GP with wavelet transform, seasonality characteristics of a complex nonlinear process could be captured. However, as also reiterated by Papacharalampous *et al.* (2017), it must be noted here that it is the specific context, available data and the process under which a model is developed that render it effective or ineffective. For example, in the present research, WGP is seen forecasting next-day Kosi flows at Baltara satisfactorily and compares well to GP and AR models, but it must be remembered that we could include monsoon data for only 13 years with limited number of high flows. This data length may be insufficient for modeling a large and complex river system like the Kosi. Moreover, during monsoon, heavy downpours in the upper reaches of Tibet and Nepal generate flooding in the Kosi, while in winter, glacier melt appears as stream flow. Therefore, it would be imprudent to suggest that the same model would be appropriate for the other rivers as well because the relevant parameters may be different.

CONCLUSION

The main objective of the present research is to examine the usefulness of the conjunction model of DWT and GP for flood forecasting in large rivers when the only data available are historical flow records. The results clearly suggest that wavelet-transfomed data improve the approximation ability of GP. The best performing WGP model, with four previous daily discharges as input, predicts the flows with the highest accuracy of 87.9% for the verification dataset of the Kosi River at the Baltara site, while autoregressive and stand-alone GP models predict flows with only 72.7% and 73.1% accuracies respectively. The peak flows are also predicted more accurately by the WGP model than by AR and GP models. In conclusion, DWT is an important data preprocessing method which denoises the flow time series and improves the efficiency of a forecasting tool.

REFERENCES

Adhikary, S. K., Muttil, N. & Yilmaz, A. G. 2016 Ordinary kriging and genetic programming for spatial estimation of rainfall in the Middle Yarra River catchment, Australia. *Hydrology Research* **47**, 1182–1197.

Ashofteh, P. S., Haddad, O. B., Alashti, H. A. & Marino, M. A. 2014 Determination of irrigation allocation policy under climate change by genetic programming. *Journal of Irrigation and Drainage Engineering* **141**, 04014059.

Aytek, A. & Kisi, O. 2008 A genetic programming approach to suspended sediment modeling. *Journal of Hydrology* **351**, 288–298.

Aytek, A., Kisi, O. & Guven, A. 2014 A genetic programming technique for lake level modeling. *Hydrology Research* **45**, 529–539.

Azamathulla, H. M., Ghani, A. A., Zakaria, N. A. & Guven, A. 2010 Genetic programming to predict bridge pier scour. *Journal of Hydraulic Engineering* **136**, 165–169.

Babovic, V. & Keijzer, M. 2002 Rainfall runoff modeling based on genetic programming. *Nordic Hydrology* **33**, 331–343.

Banzhaf, W., Nordin, P., Keller, P. E. & Francone, F. D. 1998 *Genetic Programming*. Morgan Kaufmann, San Francisco, USA.

Barge, J. & Sharif, H. 2016 An ensemble empirical mode decomposition, self-organizing map, and linear genetic programming approach for forecasting river streamflow. *Water* **8** (6), 247.

Cannas, B., Fanni, A., See, L. & Siasa, G. 2006 Data preprocessing for river flow forecasting using neural networks: wavelet transforms and data partitioning. *Physics and Chemistry of the Earth* **31**, 1164–1171.

Dorado, J., Rabunal, J. R., Pazos, A., Rivero, D., Santos, A. & Puertas, J. 2003 Prediction and modeling of the rainfall-runoff transformation of a typical urban basin using ann and gp. *Applied Artificial Intelligence: An International Journal* **17** (4), 329–343. DOI: 10.1080/713827142.

Fallah-Mehdipour, E., Bozorg Haddad, O. & Marino, M. A. 2012 Real-time operation of reservoir system by genetic programming. *Water Resources Management* **26**, 4091–4103.

Fallah-Mehdipour, E., Bozorg Haddad, O. & Marino, M. A. 2013 Developing reservoir operational decision rule by genetic programming. *Journal of Hydroinformatics* **15**, 103–119.

FMIS 2016 *Flood Management Information System*. Water Resources Department, Government of Bihar, India

Garg, V. 2014 Modeling catchment sediment yield: a genetic programming approach. *Natural Hazards* **70** (1), 39–50.

Gorgij, A. D., Kisi, O. & Moghaddam, A. A. 2017 Groundwater budget forecasting, using hybrid wavelet-ANN-GP modelling: a case study of Azarshahr Plain, East Azerbaijan, Iran. *Hydrology Research* **48**, 455–467.

Hakimzadeh, H., Nourani, V. & Amini, A. B. 2014 Genetic programming simulation of dam breach hydrograph and peak outflow discharge. *Journal of Hydrologic Engineering* **19**, 757–768.

Karimi, S., Shiri, J., Kisi, O. & Shiri, A. A. 2016 Short-term and long-term streamflow prediction by using 'wavelet-gene expression' programming approach. *ISH Journal of Hydraulic Engineering* **22** (2), 148–162.

Karimi, S., Shiri, J., Kisi, O. & Xu, T. 2017 Forecasting daily streamflow values: assessing heuristic models. *Hydrology Research* **49** (3), 658–669. DOI:10.2166/nh.2017.111.

Kisi, O. & Shiri, J. 2011 Precipitation forecasting using wavelet-genetic programming and wavelet-neuro-fuzzy conjunction models. *Water Resources Management* **25**, 3135–3152.

Kisi, O. & Shiri, J. 2012 Wavelet and neuro-fuzzy conjunction model for predicting water table depth fluctuations. *Hydrology Research* **43**, 286–300.

Kisi, O., Shiri, J. & Tombul, M. 2013 Modeling rainfall-runoff process using soft computing techniques. *Computers & Geosciences* **51**, 108–117.

Koza, J. R. 1992 *Genetic Programming: on the Programming of Computers by Means of Natural Selection*. MIT, Cambridge, MA, USA.

Li, X., Yao, J., Li, Y., Zhang, Q. & Xu, C. Y. 2016 A modeling study of the influences of Yangtze River and local catchment on the development of floods in Poyang Lake, China. *Hydrology Research* **47** (S1), 102–119.

Mallat, S. G. 1989 A theory for multi resolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**, 674–693.

Nourani, V., Baghanam, A. H., Adamowski, J. & Kisi, O. 2008 Applications of hybrid wavelet–artificial intelligence models in hydrology: a review. *Journal of Hydrology* **514** (6), 358–377.

Nourani, V., Komasi, M. & Alami, M. T. 2012 Hybrid wavelet–genetic programming approach to optimize ann modeling of rainfall-runoff process. *Journal of Hydrologic Engineering* **17**, 724–741.

Papacharalampous, G. A., Tyralis, H. & Koutsoyiannis, D. 2017 Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes. *Preprints* **2017**, 2017100133. doi: 10.20944/preprints201710.0133.v1.

Partal, T. & Kisi, O. 2007 Wavelet and neuro-fuzzy conjunction model for precipitation forecasting. *Journal of Hydrology* **342**, 199–212. doi:10.1016/j.jhydrol.2007.05.026.

Partal, T. & Kucuk, M. 2006 Long-term trend analysis using discrete wavelet components of annual precipitations measurements in Marmara region (Turkey). *Phys. Chem. Earth* **31**, 1189–1200.

Rodríguez-Vázquez, K., Arganis-Juárez, M. L., Cruickshank-Villanueva, C. & Domínguez-Mora, R. 2012 Rainfall–runoff modelling using genetic programming. *Journal of Hydroinformatics* **14**, 108–121.

Savic, D. A., Walters, G. A. & Davidson, J. 1999 A genetic programming approach to rainfall-runoff modeling. *Water Resources Management* **13**, 219–231.

Searson, D. P., Leahy, E. D. E. & Willis, M. J. 2010 *GPTIPS:An Open Source Genetic Programming Toolbox For Multigene Symbolic Regression*. gptips.sourceforge.net (accessed 15 April 2017).

Shiri, J. & Kisi, O. 2010 Short-term and long-term streamflow forecasting using a wavelet and neuro-fuzzy conjunction model. *Journal of Hydrology* **394**, 486–493.

Shiri, J. & Kisi, O. 2012 Estimation of daily suspended sediment load by using wavelet conjunction models. *Journal of Hydrologic Engineering* **17**, 986–1000.

Shiri, J., Kisi, O., Landeras, G., Lopez, J. J., Nazemi, A. H. & Louis, C. P. M. 2012a Daily reference evapotranspiration modeling by using genetic programming approach in the Basque Country (Northern Spain). *Journal of Hydrology* **414–415**, 302–316.

Shiri, J., Kisi, O., Makarynskyy, O., Shiri, A. A. & Nikoofar, B. 2012b Forecasting daily stream flow using artificial intelligence approaches. *ISH Journal of Hydraulic Engineering* **18**, 204–214.

Tiwari, K. M. & Chatterjee, C. 2011 A new wavelet-bootstrap-ANN hybrid model for daily discharge forecasting. *Journal of Hydroinformatics* **13**, 500–519.

Uyumaz, A., Mehr, A. D., Kahya, E. & Erdem, H. 2014 Rectangular side weirs discharge coefficient estimation in circular channels using linear genetic programming approach. *Journal of Hydroinformatics* **16**, 1318–1330.

Wang, W. & Ding, J. 2003 Wavelet network model and its application to the prediction of the hydrology. *Nature and Science* **1**, 67–71.

Yaseen, Z. M., El-shafie, A., Jaafar, O., Afan, H. A. & Sayl, K. N. 2015 Artificial intelligence based models for stream-flow forecasting: 2000–2015. *Journal of Hydrology* **530**, 829–844.

First received 13 November 2017; accepted in revised form 25 March 2018. Available online 23 April 2018