Short-term flood forecasting with an amended semi-parametric regression ensemble model

Le Hoang Tuan, and To Anh Dung

***Abstract*—Flood forecasting is very important research topic in disaster prevention and reduction. The characteristics of flood involve a rather complex systematic dynamic under the influence of different meteorological factors including linear and non-linear patterns. Recently there are many novel forecasting methods of improving the forecasting accuracy. This paper explores the potential and effect of the semi-parametric regression to modelize flood water-level and to forecast the inundation of Mekong Delta in Vietnam. The semi-parametric regression technique is a combination of a parametric regression approach and a non-parametric regression concept. In the process of model building, three altered linear regression models are applied for the parametric component. They are stepwise multiple linear regression, partial least squares solution and multi-recursive regression method. They are used to capture flood's linear characteristics. The non-parametric part is solved by a modified estimation of a smooth function. Furthermore, some justified non-linear regression models based on artificial neural network are also able to obtain flood's non-linear characteristics. They help us to smooth the model's non-parametric constituent easily and quickly. The last element is the model's error. Then the semi-parametric regression is used for ensemble model based on the principle component analysis technique. Flood water-level forecasting, with a lead time of one and more days, has been made by using a selected sequence of past water-level values and some relevant factors observed at a specific location. Time-series analytical method is utilized to build the model.**

Obtained empirical results indicate that the prediction by using the amended semi-parametric regression ensemble model is generally better than those obtained by using the other models presented in this study in terms of the same evaluation measurements. Our findings reveal that the estimation power of the modern statistical model is reliable and auspicious. The proposed model here can be used as a promising alternative forecasting tool for flood to achieve better forecasting accuracy and to optimize prediction quality further.

***Index Terms*—parametric, non-parametric, semi-parametric regression ensemble model, stepwise multiple linear, partial least squares, multi-recursive regression, artificial neural network, estimation, smooth, kernel function, splines, flood, water-level, prediction, forecasting**.

Manuscript Received on July 13th, 2016. Manuscript Revised December 06th, 2016.

The authors thank the University of Information Technology – Vietnam National University Ho Chi Minh City, Vietnam, for your help in our research. This research is funded by University of Information Technology - Vietnam National University Ho Chi Minh City under grant number D1-2017-05.

Le Hoang Tuan is with the Department of Mathematics and Physics, University of Information Technology - Vietnam National University Ho Chi Minh City, KM20, Hanoi Highway, Block 6, Linh Trung Ward, Thu Duc Dist., Ho Chi Minh City, Vietnam (e-mail: tuanlh@uit.edu.vn).

To Anh Dung was with the Faculty of Mathematics and Computer Science - Vietnam National University Ho Chi Minh City, 227 Nguyen Van Cu St., Dist. 5, Ho Chi Minh City, Vietnam (e-mail: tadung@hcmus.edu.vn).

1 INTRODUCTION

Flood prediction is a challenging task in climate dynamics and climate conjecture theory. Accurate and timely flood prognostication is important and essential for the planning, management, and development of water resources, in particular for spate warning systems. It can provide an information to help preventing casualties and damages caused by natural calamities [10]. For instance, a deluge warning system for fast responding catchments may require a quantitative flood forecast to increase the lead time for warning. Additionally, freshet prediction is one of the most complex and difficult elements of the hydrology cycle to understand and to modelize due to the complexity of involved atmospheric processes and the variability of inundation in space and time. Especially, Vietnam is a tropical and monsoon country, with high rates of rainfall and humidity that are affected by climate change. Floods happen more and more with an increasing frequency and devastation. To help people to subsist on floods, to reduce human and material losses to the minimum are the chief target of our

society. Short-term flood prediction is one of effective solutions for this problem.

Traditionally, autoregressive moving average (ARMA) has been used in modelling and forecasting water resource time series because such models are accepted as a standard representation of a stochastic time series [*Maier and Danny, 1997*]. The method that is based on a statistical technique makes use of classical statistics to analyze historical data with an objective to evolve methods for the formulation of flood forecasts [e.g., *Box and Jenkins*, 1976; *Salas and Obeysekera*, 1982; *Sharma*, 1985]. However, such models do not attempt to represent the non-linear dynamics inherent in the transformation of rainfall to runoff and therefore may not always perform well [*Hsu et al.*, 1995]. Owing to the difficulties associated with non-linear model structure identification and parameter estimation, very few exactly non-linear system theoretic watershed models have been reported [*Jacoby*, 1966; *Amorocho and Brandstetter*, 1971; *Ikeda et al.*, 1976]. In most cases, linearity or piecewise linearity has been assumed [*Hsu et al.*, 1995]. Therefore it seems necessary that the conventionally applied modelling solutions be refined or supplemented to get improved performance by implementing new or different technologies.

To date, a plethora of rainfall-runoff models belonging to different categories is available for flood forecasting purposes. They include conceptual models that try to conceptualize many physical processes influencing the runoff, empirical models, and complex models that couple meteorologic and hydrologic patterns for flow forecasting. In the physical approach, the primary motivation is the study of physical phenomena and their understanding, while in the system theoretic accession the concern is with system's operations, not the nature of the system by itself or the physical law governing its operation [3]. Besides, we have several modern softwares which used in flood water-level forecasting, such as: MARINE, SSARR, TANK, NAM, MIKE11, DIMOSOP, HYDROGIS,…Most of the applied configurations are one dimensional hydrologic, hydraulic or hydro-dynamic models, that utilize St. Venant adequately simultaneous equations. They are the most popular flood forecasting models.

Nevertheless, the concept of coupling different models has been a widely accepted research topic in hydrologic forecasting, which has attracted scientists from other fields including Statistics, Machine Learning, and so on [11]. They can be broadly categorized into ensemble models and modular (or hybrid) forms. The principal idea behind ensemble models is to establish severally different or similar models for the same process and to integrate them together. Their success largely arises from the fact that they lead to improved accuracy that is compared to a single classification or a regression model. Typically, ensemble techniques comprise two phase: a) the production of multiple predictive models, and b) their combination. Recent work has been the main consideration the reduction of ensemble sizes prior to combination.

Recently, more hybrid forecasting approaches have been advanced to upgrade flood prediction accuracy and rapidity. Ashu Jain et al [4] have applied Hybrid neural network models in hydrologic time series forecasting. The proposed technique consists of an overall modelling framework, which is a conjunction of four time series models of auto-regressive type and four artificial neural network models. The obtained results in this study advise that the proposal of combining some strengths of conventional and artificial neural network performances dispenses a robust modelling framework capable of capturing the non-linear nature of multiple complicated time series and thus propagating authentic forecasts. Some lately investigations clearly expose that many hybrid models could equip an effective way to optimize forecasting accuracy and celerity that are achieved by either of the models used separately [6].

Although it is easy for someone to find many applications of regression ensemble models in a variety of areas, such attempts have been limited in the model of flood forecasting. Unfortunately, there are hybrid linear methods or integrated non-linear techniques for flood forecasting. Some probable reasons for arduousness in prognosticating inundation are the complexity of atmosphere – ocean interactions and the uncertainty of the relationship between flood and hydro-meteorological variables. The characteristic of deluge involves a rather complex system dynamics under the influence of numerous meteorological factors. They often contain both linear and non-linear patterns so that this paper presents and demonstrates the applicability of an effective semi-parametric regression ensemble model, which is coupled with appropriate data-preprocessing techniques by justified parametric estimations and

suggested non-parametric solutions to upgrade the accuracy and velocity of short-term flood forecasting process. The daily and hourly flood water-level values at a given streamflow gauge station with different lead times, have been predicted in Long Xuyen quadrangular basin (a specific zone of the Lower Mekong Delta) in Vietnam. The paper also aims at an extensive evaluation of a semi-parametric regression ensemble model with pure parametric estimations and purified non-parametric solutions based on a model in short-term flood forecasting, and critical comments on their relative merits and limitations.

2 Case Study

2.1 Study area

The measured flood water-level data were available at the Chau Doc, Long Xuyen, Tri Ton, Xuan To, and Tan Chau gauging stations, in An Giang province, Vietnam. Tan Chau station is coded by 019803, located on upstream of Tien River, at longitude 105º13'' and lattitude 10º45'. Chau Doc station is coded by 039801, located on upstream of Hau River. These stations are settled in Long Xuyen quadrangular basin, one of areas sustained heavy losses in the inundations in Mekong Delta every year. It is shown in Fig. 1.

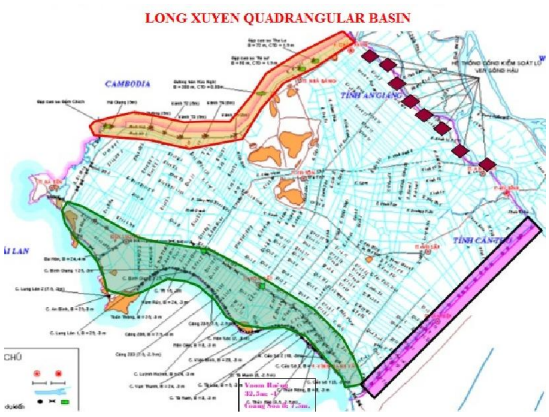

Figure 1. Catchment area plan

This catchment is approximately 489000 hectare natural area. The topography is sunken, even and flat with nearly from 0.4 (m) to 2.0 (m) altitude from the sea water level. Yearly, the flood season occurs from June to November, especially highest in August and September. This studied basin is often inundated from 0.5 (m) to 2.5 (m) depth, although there is a dense man-made canals, embankments and road systems. Flood is moderate but exists in long duration. If deluges arise, entire North part will be immersed. Flooding is caused by tidal influence, overlanding flow border (with approximately eighty percent in the main streams). Then the irregular change of upstream head-waters of Mekong River, especially from the border between Cambodia and Vietnam, could lead into fluctuations.

2.2 The data

Daily 24-hours flood water-level values in fifteen years, from 1st January, 2000 to 31st December, 2014, were extracted from the weekly reports of the Regional Flood Management and Mitigation Centre, a division of Mekong River Commission. Besides, meteorological data at the same periods were also collected at several hydrographic stations. They include water-level, precipitation, evaporation, air-humidity, ground-moisture, wing-speeds, air-directions values, and so on. Then the data were divided into two separate groups randomly. The first group was used for training, and the second one was applied for testing and calibration. While model building, every seven successive days is gathered to establish sub-groups. In these groups, the five first daily flood values are the input elements and two remaining ones are the output constituents. Thus, in all, 87840 input-output data records were used successively for training and 43920 data records were used for testing application.

The objective is to model and to forecast daily flood water-level values with lead time of 1 and 2 days. Since the main purpose of this paper is to furnish citizens with short-term forecasted results, we do not carry out the algorithms for 3-days, 4-days and beyond. The final received results from the amended semi-parametric regression ensemble model, via a dimension-reduction subspace algorithm, justified parametric estimations and altered non-parametric solutions, could be helpful basic information for model adjusting, extending and upgrading. In other words, even though a larger lead time of model or forecast would be more useful to issues the flood warnings well in advance, a smaller lead time can help in making emergency reservoir operations and in cautioning the population at longer distances downstream or at many specific sites where a nearby river gauging station is not available.

3 The Building Process Of The Semi-parametric Regression Ensemble Model

The most popular mathematical model for

making predictions is the multiple linear regression model in the traditional mathematical technique, which is an estimation of one random variables by using known values of multiple variables. There is a continuous random variable called the dependent variable, Y ; there are many independent variables, $X_1, X_2, ..., X_p$. The target of regression models is to fit a set of data with an equation, the simplest being a linear equation. The linear regression model is given by

$$\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon \quad (1)$$

where ε , the "noise" variable factor, is a normally distributed random variable with mean equal to zero and standard deviation σ , $\{\beta_t, t = 1,2,...,p\}$ are some values of the coefficients. Unknown parameters values should be estimated from the smallest error sum of squares of the samples.

A habitual obstacle in modelling is that of use of a regression equation to predict the value of a dependent variable when the actual data have a number of variables to choose at independent variables in model. The choice of appropriate elements is very important for a complete model. Estimates of regression coefficients are likely to be unstable due to multi-colinearity in models with many independent variables. Before conjecturing relative parameters, the effective dimension-reduction subspace algorithm which was proposed by Arnak S. Dalalyan et al. [7] is used to eliminate some factors that slightly affect to a model. Then three justified linear regression models have been made in this paper for parsimony variables. They are stepwise multiple linear (SML) regression, partial least square (PLS) regression, and multi-recursive regression (MR) methods. They are tried to different significant levels, to capture the linear characteristics of inundations.

Furthermore, we also execute non-parametric estimations via different methods. They involve ameliorated kernel smooth functions, modular spline techniques, and three modified algorithms in artificial neural network. They provide an interesting solution that theoretically can approximate any non-linear continuous function on a compact domain to any designed accuracy and velocity [4]. Several non-parametric estimations are employded in order to supply a plenty of efficient anwers for choosing the best suitable non-linear solution.

The artificial neural network is chosen in this paper because it learns by adjusting the interconnections among layers flexibility. When a network is adequately trained, it is able to learn a relevant output for a set of input data. There are some advantages when using a neural network for flood water-level modeling and forecasting, such as: neural network is designed to recognize a hidden pattern in data in a similar way to that of the human brain, neural network is useful when the underlying problem is either poorly defined or not clearly understood, its applications do not requrire a prerequisite knowledge about the studied process, and so on [12]. In the practical application, many results of experiments have shown that the generalization of single neural network is not unique. That is, artificial neural network results are not stable. Even for some simple problems, different structures of neural networks (e.g., different numbers of hidden layers, multiple hidden nodes and numerous initial conditions) result in different patterns of network generalization. If carelessly used, it can easily lead to irrelevant information in many systems and limited applications of artificial neural network in the flood forecasting [10]. Due to the work about bias-variance trade-off of by Bretiman [8], an artificial neural network regression model consisting of diverse models with much disagreement is more likely to have a good generalization [10].

In this paper, three justified neural network algorithms that were suggested in [12] are applied together with several mention-aboved non-parametric estimations to obtain flood's non-linear characteristics.

3.1 Semi-parametric regression model

In recent years, semi-parametric regression model is a popular accepted regression model which has been widely applied to many fields such as economics, medical science and so on [2]. It is an emerging field that represents a fusion between traditional parametric regression analysis and newer non-parametric regression techniques. It synthesizes research across several branches of statistics: parametric and non-parametric regression, longitudinal and spatial data analysis, mixed, etc. It is also a deeply rooted field in applications and its evolution reflects the increasingly large and complex problems that are arising in science and industry [5].

Parametric regression technique which realized the pure parametric thinking in curve estimations often does not meet the need in complicated data analysis. Some alternatives are highly flexible non-parametric regression solutions, the object of which is to estimate a regression function directly, rather

than to estimate parameters. Nonetheless, in many applications it is necessary to come up with a decision whether a covariant is essential for understanding of the problem or not. Hence several kinds of parameter testing are required, which can not be performed in a purely non-parametric setting. An alternative solution is a semi-parametric regression model with a predictor fuction consisting of a parametric linear component and a non-parametric element which involves some additional predictor variables. Semi-parametric regression can be of worthwhile value in the solution of complex scientific difficulties.

Suppose our data consist of n subjects. For subject ($k = 1,2,...,n$), Y_i is the independent variable, X_i is a m vector of clinical covariates which we mentioned above, and $Z_i = X_i^T \theta_0$ is a p vector of gene expressions within a pathway. The outcome Y_i depends on both X_i and Z_i. Besides, β_0 is a m vector of regression coefficients, $g(Z_i)$ is an unknown centered smooth function, and the errors ε_i are assumed to be independent and followed $N(0,\sigma^2)$. Then the proposed semi-parametric regression model is given by

$$Y_i = X_i^T \beta_0 + g(X_i^T \theta_0) + \varepsilon_i = X_i^T \beta_0 + g(Z_i) + \varepsilon_i \quad (2)$$

where $X_i^T \beta_0$ is the parametrical part of model for epitaxial forecasting. Its target is to control a independent variable trend. Here $g(Z_i)$ is the non-parametrical part of model for local calibration so that it is better to fit responses value. So model includes the effects of parametrical element and the profits of non-parametrical part.

A solution can be achieved by minimizing the sum of squares equation

$$J(g,\beta_0) = \sum_{i=1}^{n} (y_i - x_i^T \beta_0 - g(z_i))^2 + \lambda \int_a^b [g''(z_i)]^2 dt \quad (3)$$

where $\lambda \geq 0$, is a tuning parameter which controls the tradeoff between goodness of fitting and complexity of our model. When $\lambda = 0$, the model interpolates the gene expression data, whereas, when $\lambda = \infty$, the model reduces to a simple linear model without $g(.)$ [10]. Based on earlier works, especially many suggested iterative procedures in [10], the resulting estimator is often called a partial spline. It is known because this estimator is assymtotically biased for the optimal λ choice when all components of β depend on t. An asymtotic bias can be larger than a standard error. Correlation between predictors is common in real data analysis.

Additionally, several kernel smooth functions are applied to solve the non-parametric estimations. They are listed in the Table 1.

3.2 The semi-parametric regression ensemble

Flood forecasting problem is far from simple due to water-level, precipitation, evaporation, air-humidity, ground-moisture, etc. with high complexity, irregularity and noisy. In this paper, three altered linear regressions are used to capture flood's linear characteristics, while the rest of ameliorated solutions are capable of clutching non-linear patterns in weather system. Ensemble the eight models may yield a robust method, and more satisfactory forecasting results may be obtained by combining linear regression part and non-linear regression model.

TABLE 1
KERNEL FUNCTIONS

Kernel	$K(u)$
Uniform/ Boxcar	$\frac{1}{2} I(\lvert u \rvert \leq 1)$
Triangle	$(1-\lvert u \rvert) I(\lvert u \rvert \leq 1)$
Epanechnikov	$\frac{3}{4}(1-u^2) I(\lvert u \rvert \leq 1)$
Quartic (Biweight)	$\frac{15}{16}(1-u^2)^2 I(\lvert u \rvert \leq 1)$
Triweight	$\frac{35}{32}(1-u^2)^3 I(\lvert u \rvert \leq 1)$
Gaussian	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right)$
Cosine	$\frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right) I(\lvert u \rvert \leq 1)$
Tricube	$\frac{70}{81}(1-u^3)^3 I(\lvert u \rvert \leq 1)$

Let $\hat{y}_1, \hat{y}_2$, and $\hat{y}_3$, respectively, be the forecasting output of linear regression models, while $\hat{y}_4$ be the output of kernel smooth estimation and $\hat{y}_5$ be the output of spline technique. Let $\hat{y}_6, \hat{y}_7$, and $\hat{y}_8$, respectively, be the non-linear output of three modified artificial neural network models. Then the Principle Component Analysis (PCA) technique is utilized to extract an effective feature from all forecasting output matrix. So the semi-parametric regression ensemble (SRE) model

is established.

The above-mentioned solution can be summed up as follows: firstly, three different linear regression models are used to get linear forecasting outputs. Secondly, a plenty of algorithms training non-parametric estimations are used to get non-linear prediction outputs. Thirdly, the PCA technique extracts ensemble members from linear and non-linear forecasting outputs. Finally, SRE is used to combine the selected individual forecasting results into a semi-parametric ensemble model.

These ideas are shown by the following diagram in Fig. 2.

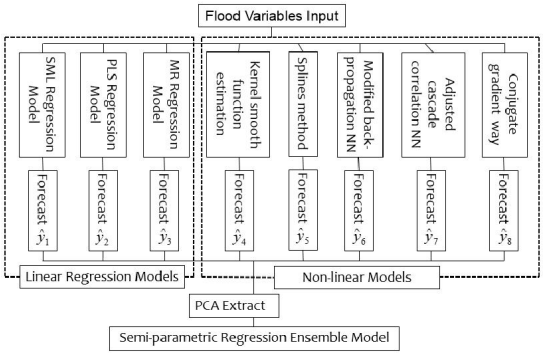

Figure 2. A flow diagram of the proposed semi-parametric regression ensemble model

For the purpose of comparision, the paper has also established other three ensemble forecasting models. They are given by:

(1) simple average all the linear (SAL) forecasting outputs

$$\hat{Y}_1 = \hat{y}_1 + \hat{y}_2 + \hat{y}_3 \quad (4)$$

(2) simple average all the non-linear (SAN) forecasting outputs

$$\hat{Y}_2 = \hat{y}_4 + \hat{y}_5 + \hat{y}_6 + \hat{y}_7 + \hat{y}_8 \quad (5)$$

(3) stepwise regression all the linear and non-linear (SRLN) forecasting outputs

$$\hat{Y}_3 = w_1\hat{y}_1 + w_2\hat{y}_2 + w_3\hat{y}_3 + w_4\hat{y}_4 + w_5\hat{y}_5 + w_6\hat{y}_6 + w_7\hat{y}_7 + w_8\hat{y}_8 \quad (6)$$

where the weights of factors can be solved according to (1) in Section 2.

Additionally, the credible intervals with different significant levels are proposed for testing purpose. They are built up relying on interval estimations of parametric and non-parametric regression models in [2], [5]. They are one of criteria to evaluate the precision and suitability of a model. They are defined by

$$\left[\bar{y}_{predict} - t_{\frac{\alpha}{2}}\left(\frac{S}{\sqrt{n}}\right), \bar{y}_{predict} + t_{\frac{\alpha}{2}}\left(\frac{S}{\sqrt{n}}\right)\right] \quad (7)$$, where $\bar{y}_{predict}$ is a forecasting output of the final semi-parametric regression ensemble model, while $t_{\frac{\alpha}{2}}$ is an extra value that is derived from the Gaussian distribution Appendix table in [2], [5], corresponding to a significant level α, and S is a justified standard deviation of the model's sample.

The velocity of the model convergence is also investigated.

4 Empirical Analysis And Results

Real-time flood data are obtained in fifteen years, from 1st January, 2000 to 31st December, 2014, in the Long Xuyen quadrangular basin, by observing six gauging stations. 87840 sample data records are uttilized to build the model, other 43920 data records are tested in modelling process.

Method of modelling is one-step ahead prediction, that is, the forecast is only one sample each time and the training sample is an additional one each time on the basis of the last training. A laptop, with CPU Intel core i5 processor, is used to grant for this research.

First of all, some candidate forecasting factors are selected from the hydrological-meteorological data, which includes: water-level, precipitation, evaporation, air-humidity, ground-moisture, and other meteorological/ physical elements from many weekly reports of the Regional Flood Management and Mitigation Centre, a division of Mekong River Commission. We can get twenty variables as the main forecasting factors. The original data are used as real outputs.

The configuration of some artificial neural network models, the number of iterations to achive an overall mean square error of the 1 (cm), are given in Table 2 and Table 3 for warning time of 1 and 2 days.

TABLE 2
Training Details For Altered Back-propagation Algorithm

Year	Network configuration			Iterations	Time (*s*)
	I	O	H		
2009	5	3	2	35100	1053
2010	5	3	2	52920	1587.6
2011	5	3	2	48600	1458
2012	5	3	2	45260	1340
2013	5	3	2	42170	1224
2014	5	3	2	40480	1150

TABLE 3
TRAINING DETAILS FOR MODIFIED CASCADE CORRELATION ALGORITHM

Year	Network configuration			Iterations	Time (*s*)
	I	O	H		
2009	6	2	2	8775	263.25
2010	6	2	2	11760	352.80
2011	6	2	2	10125	303.75
2012	6	2	2	9862	295.75
2013	6	2	2	8625	282.50
2014	6	2	2	7860	275.25

Besides, maximum error (ME1), minimum error (ME2), average value of errors (AE), normalized maximum and minimum values (ME3 and ME4), maximum and minimum values of η (ME5, ME6) and α (ME7, ME8) are also given in Table 4.

TABLE 4
SOME SPECIFIC VALUES OF THE JUSTIFIED BACK-PROPAGATION ALGORITHM

Year	2009	2010	2011
ME1	1.3200	1.1500	1.0202
ME2	2.5E-16	2.4E-16	2.1E-18
AE	2.8576E-8	2.1679E-8	2.1554E-8
ME3	0.9738	0.7500	0.9905
ME4	0.1976	-0.0500	0.0905
ME5	1.0146	0.9999	0.9804
ME6	0.5513	0.6096	0.6328
ME7	0.9999	1.9300	0.9900
ME8	0.9900	0.8020	0.9000

In order to measure an effectiveness of the proposed model, three types of errors, such as, normalized mean squared error (NMSE), mean absolute percentage error (MAPE) and Pearson Relative Coeeficient (PRC) error, which have been found in many papers, are also utilized here. Interested readers can be refered to [1] for more details.

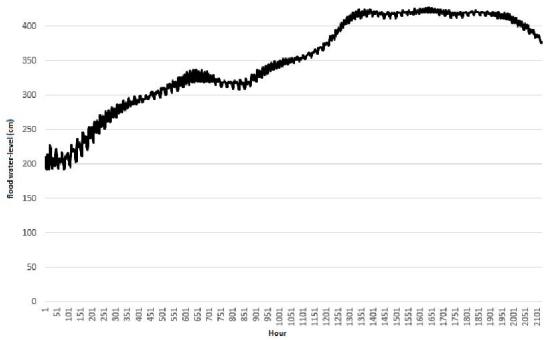

Figure 3. Fitting results for training samples

Fig. 3 gives a graphical representation of many fitting results for flood water-levels of all regions by using different models, which are used to fit flood water-level values of training samples for comparision.

Moreover, Table 5 illustrates the fitting accuracy and efficiency of our model in terms of various evaluation indices for training samples.

TABLE 5
A COMPARISION OF RESULT OF ENSEMBLE MODELS ABOUT 2103 TRAINING SAMPLES

Results	SAL	SAN	SRLN	SRE
NMSE	32.2508	21.4825	20.3618	12.1453
MAPE (%)	24.6815	18.1564	16.4523	10.1625
PRC	0.6812	0.7856	0.8863	0.9214

Fig. 4 shows forecasting of testing samples by using different models, which are used to predict the flood water-level values of 701 testing samples for comparision.

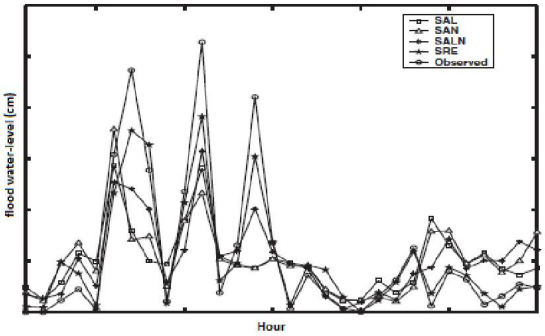

Figure 4. Testing results for 701 testing samples

Many received results reveal that learning ability of a semi-parametric regression ensemble outperforms the other models under the same input conditions. Another important factor to measure performance of a method is to check its forecasting ability of testing samples for actual inundation events.

Table 6 shows the forecasting performance of various modes from different perspectives in terms of plentiful evaluation indices. From the graphs and tables, we can generally see that the forecasting results are very promising for all flood regions under study either where the measurement of fitting performance is goodness of fit such as NMSE or where the forecasting performance effective is PRC (refer Table 5 and Table 6).

TABLE 6
A COMPARISION OF RESULT OF ENSEMBLE MODELS ABOUT 701 TRAINING SAMPLES

Results	SAL	SAN	SRLN	SRE
NMSE	36.8315	27.4536	24.3852	13.6175
MAPE (%)	25.5327	21.1834	19.1625	11.5324
PRC	0.5862	0.7568	0.7952	0.9014

On the other hand, from the presented experiments in this study, we can draw that the semi-parametric regression ensemble model is superior to other models about the fitting and testing cases in terms of different measurements. There are many reasons for this phenomenon. Firstly, the inundations contain complex linear and non-linear patterns, the proposed model can extract sophisticated trends and can find many structures in flood time series. Using different kernel function forms of an effective non-linear mapping, together with spline algorithms and several modified artificial neural network solutions, we can establish several effective and suitable non-linear mappings for flood forecasting. Secondly, the output of different models has a high correlative relationship, a high noise, non-linearity and complex factors. If the suggested technology does not reduce the dimension of data and does not extract main features, the results of a model is unstable. At last, SRE is used to combine some selected individual flood forecasting results into a semi-parametric regression ensemble model, which keeps the flexibility of both linear and non-linear models. So the proposed semi-parametric regression ensemble model can be utilized as a feasible approach to short-term flood forecasting.

5 CONCLUSION

Accurate flood forecasting is crucial for a frequent unanticipated flood region to avoid life losing and economic loses. This study proposes using a semi-parametric regression ensemble forecasting model that combines a linear regression part and a non-linear element to predict flood based on PCA technique. In terms of empirical results, we find that across numerous forecasting models for the short-term forecasting samples of regions in Long Xuyen quadrangular basin, on the basis of different criteria, the amened semi-parametric regression ensemble model performs the best.

In the process of testing samples of the proposed semi-parametric regression ensemble model, the NMSE is the lowest and the PRC is the highest, indicating that the suggested regression ensemble forecasting model can be accepted as a reliable alternative solution for short-term flood forecasting. Moreover, a semi-parametric regression ensemble model can provide more reference information for future prediction, has no unreliable information. Those results indicate that the proposed model here can be applied as a promising alternative forecasting tool for flood to achieve better forecasting accuracy and to optimize prediction quality and velocity further. The model should be continued to develop and apply in our real world, emphasized in other regions that have some similar characteristics to the studied area.

REFERENCES

[1] Jiansheng Wu, Mingzhe Liu, and Long Jin, "Least square support vector machine ensemble for daily rainfall forecasting based on linear and nonlinear regression," *Advanced in Neural Network Research & Application*, Lecture Notes in Electrical Engineering, the Springer Press, vol. 12, no. 10, pp. 993-1001, 1990.

[2] W. Härdle, M. Müller, S. Sperlich, and A. Werwatz, "Semiparametric regression. Nonparametric and semiparametric model," the Springer Press, 2004.

[3] P. C. Nayak, K. P. Sudheer, D. M. Rangan, and K. S. Ramasastri, "Short-term flood forecasting with a neurofuzzy model," *Water Resourses Research*, vol. 41, W04004, 2005, DOI. 10.1029/2004WR003562.

[4] J. Ashu, and M. K. Avadhnam, "Hybrid neural network models for hydrologic time series forecasting," *Applied Soft Computing*, vol. 7, pp. 585-592, 2007.

[5] D. Rupper, M. P. Wand, and R. J. Carroll, "Semiparametric regression," durging 2003-2007, [Online] Available: http://uow.edu.au/mwand/sprpap.pdf.

[6] W. C. Hong, "Rainfall forecasting by technological machine learning model," *Application of Mathematics and Computation*, vol. 200, pp. 41-57, 2008.

[7] Arnak S. Dalalyan, Anatoly Juditsky, and Vladimir Spokoiny, "A New Algorithm for Estimating the Effective Dimension-Reduction Subspace," *Journal of Machine Learning Research*, vol. 9, pp. 1647-1678, 2008.

[8] L. N. Yu, S. Y. Wang, and K. K. Lai, "Neural network based mean-variance-skewness model for portfolio selection," *Computer and Operations Research*, vol. 35, pp. 34-46, 2008.

[9] J. Wu, "A Semi-parametric Regression Ensemble Model for Rainfall Forecasting Based on RBF Neural Network," *Wang, F. L., Deng, H., Gao, Y., Lei, J. (eds.) AICI 2010, Part II. LNCS (LNAI)*, vol. 6320, pp. 284-292, 2010, Springer-Heidelberg (2010).

[10] Jiansheng Wu, "An Effective Hybrid Semi-Parametric Regression Strategy for Artificial Neural Network Ensemble and Its Application Rainfall Forecasting," *Fourth International Joint Conference on Computational Sciences and Optimization*, pp. 1324-1328, 2011, 978-0-7695-4335-2/11, DOI. 10.1109/CSO.2011.71.

[11] J.-S. Pan, S.-M. Chen, and N. T. Nguyen, "Prediction of Rainfall Time Series Using Modular RBF Neural Network Model Coupled with SSA and PLS," *ACIIDS 2012*, Part II, LNAI 7197, pp. 509-518, 2012, Springer-Verlag Berlin Heidelberg (2012).

[12] Le Hoang Tuan, and To Anh Dung, "A Modified Semi-parametric Regression Model For Flood Forecasting," *Science and Technology Development*, vol. 18, no. K4-2015, pp. 95-105, 2015.

[13] Jianzhu Li, Senming Tan, "Nonstationary Flood Frequency Analysis for Annual Flood Peak Series," *Adopting Climate Indices and Check Dam Index as Covariates*," *Water Resource Manage*, vol. 29, pp. 5533-5550, 2015, DOI. 10.1007/s11269-015-1133-5.

Dự báo lũ lụt ngắn hạn bằng mô hình tập hợp hồi quy bán tham số có cải biên

Lê Hoàng Tuấn, Tô Anh Dũng

Tóm tắt - **Dự báo lũ lụt là chủ đề nghiên cứu đóng vai trò hết sức quan trọng trong công tác ngăn ngừa và giảm thiểu thiệt hại do thiên tai gây ra. Những tính chất đặc trưng của lũ lụt thường có liên quan đến một hệ thống động lực khá phức tạp, dưới sự ảnh hưởng và tác động của nhiều yếu tố khí tượng thủy văn khác nhau, bao gồm cả thành phần tuyến tính lẫn phi tuyến. Những năm gần đây nhiều phương pháp dự báo mới có liên quan đến việc cải tiến độ chính xác của dự báo đã được hình thành. Bài báo này tiến hành nghiên cứu tiềm năng và tính hiệu quả của mô hình hồi quy bán tham số cho việc mô hình hóa bề mặt mực nước lũ lụt và việc dự báo lũ lụt ở vùng Đồng bằng Sông Cửu Long, Việt Nam. Kỹ thuật hồi quy bán tham số là một sự kết hợp giữa cách tiếp cận hồi quy tham số và quan điểm hồi quy phi tham số. Trong quá trình xây dựng mô hình, ba mô hình hồi quy tuyến tính có hiệu chỉnh được sử dụng để giải quyết cho thành phần tham số. Đó chính là hồi quy tuyến tính bội bậc thang, bình phương bé nhất riêng phần, và phương pháp hồi quy đa đệ quy. Những giải pháp này được đưa ra nhằm giúp cho ta nắm bắt được tính chất tuyến tính của lũ lụt. Thành phần phi tham số của mô hình thì được xử lý bằng phép ước lượng cải biên từ hàm làm trơn. Ngoài ra, một vài mô hình hồi quy phi tuyến (đã qua xử lý) dựa trên nền tảng mạng nơ ron nhân tạo cũng được khảo sát đến, nhằm cung cấp cho ta thông tin về tính chất phi tuyến của lũ lụt. Chúng giúp việc làm trơn thành phần phi tham số của mô hình được diễn ra nhanh chóng và dễ dàng. Thành phần sau cùng của mô hình dự báo là sai số của mô hình. Sau đó, phương pháp phân tích thành phần chính được vận dụng để hình thành nên mô hình tập hợp hồi quy bán tham số. Dự báo bề mặt mực nước lũ lụt ngắn hạn, với khoảng thời gian một-hai ngày, được tiến hành dựa trên chuỗi dữ liệu giá trị bề mặt mực nước từ quá khứ, cùng với thông tin về những yếu tố liên quan đã ghi nhận, ở một vị trí cụ thể. Phương pháp phân tích chuỗi thời gian cũng được nghiên cứu vận dụng để xây dựng mô hình. Các kết quả thực nghiệm nhận được cho thấy rằng việc dự đoán bằng cách sử dụng mô hình tập hợp hồi quy bán tham số có cải biên nói chung là tốt hơn so với những mô hình khác được giới thiệu trong phạm vi bài báo này, xét trên cùng một đơn vị đánh giá. Những phát hiện này nói lên rằng năng lực ước lượng của mô hình thống kê hiện đại là rất đáng tin cậy, khả thi và có lợi cho ta. Mô hình được đề xuất ở đây có thể được sử dụng như là một công cụ dự báo thay thế đầy triển vọng trong tương lai, cho lĩnh vực dự báo lũ lụt, nhằm đạt được độ chính xác dự báo tốt hơn, và tối ưu thêm chất lượng tiên đoán.**

Từ khóa - **tham số, phi tham số, mô hình tập hợp hồi quy bán tham số, tuyến tính bội bậc thang, bình phương bé nhất riêng phần, hồi quy đa đệ quy, mạng nơ ron nhân tạo, ước lượng, làm trơn, hàm hạt nhân, hàm spline, lũ lụt, bề mặt mực nước, tiên đoán, dự báo.**